DOCUMENT RESUME

ED 128 442	TM 005 631
AUTHOR	Frederiksen, Norman; Ward, William C.
TITLE	Development of Measures for the Study of Creativity.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RB-75-18; GREB-72-2P
PUB DATE	Jun 75
NOTE	153p.
EDRS PRICE	MF-$0.83 HC-$8.69 Plus Postage.
DESCRIPTORS	*Behavioral Science Research; Correlation; *Creative Thinking; *Creativity Tests; Factor Analysis; Graduate Students; Higher Education; Job Analysis; *Performance Tests; *Problem Solving; Researchers; Response Style (Tests); Scoring Formulas; Simulation; Test Reliability; Test Validity
IDENTIFIERS	*Test of Scientific Thinking

ABSTRACT

A set of tests that might be reasonably used as provisional criterion measures in research on scientific thinking, particularly creative thinking, were developed and an assessment was made of the suitability of these tests as criterion variables from the standpoint of their psychometric properties. The Tests of Scientific Thinking are performance tests that simulate aspects of the job of a behavioral scientist. The tests are: Formulating Hypotheses, Evaluating Proposals, Solving Methodological Problems, and Measuring Constructs. The examinee proposes a number of solutions--not only the one considered best, but also others that should be considered. A scoring method was developed that requires the scorer to assign values to categories of responses rather than to make subjective evaluations. Six scores were studied: (1) average quality of the responses the examinee thinks are best; (2) average quality of all responses; (3) average quality of the best response by category scoring; (4) number of responses; (5) number of unusual responses; and (6) number of responses that are both unusual and of high quality. The tests were administered to about 4,000 graduate school applicants using an item sampling procedure. Test difficulty was found appropriate for advanced students and reliabilities were high enough to be useful. Factors analyses were performed to clarify the structure of the interrelationships among the various scores for the four tests. The tests seemed face valid, but evidence of construct validity is needed. (RC)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

THE GRADUATE RECORD EXAMINATIONS

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
DOROTHY URBAN
EDUCATIONAL TESTING SERVICE
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

GRE

DEVELOPMENT OF MEASURES FOR THE STUDY OF CREATIVITY

Norman Frederiksen
William C. Ward

GRE Board Professional Report GREB No. 72-2P

Research Bulletin RB-75-18
(May 1975)

June 1975

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

ETS
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY □ BERKELEY, CALIFORNIA □ EVANSTON, ILLINOIS

DEVELOPMENT OF MEASURES
FOR THE STUDY OF CREATIVITY

Norman Frederiksen
William C. Ward

GRE Board Professional Report GREB No. 72-2P

Research Bulletin RB-75-18
(May 1975)

June 1975

Copyright © 1975 by Educational Testing Service. All rights reserved.

DEVELOPMENT OF MEASURES FOR THE STUDY OF CREATIVITY

Abstract

Research on creative thinking has been handicapped by lack of adequate criteria. The purpose of this study was to develop a set of tests that could be used as dependent measures in evaluating training or in other research on "creativity." Four tests were developed, called Formulating Hypotheses, Evaluating Proposals, Solving Methodological Problems, and Measuring Constructs. They are job-sample tests that present realistic tasks such as a behavioral scientist might have to deal with. A scoring method was developed that requires the scorer to assign responses to categories of responses rather than to make subjective evaluations. The categories are assigned scale values based on independent evaluations by an expert panel. Scores can be assigned by computer. Six scores were studied: (1) average quality of the responses the examinee thinks are best, (2) average quality of all responses, (3) average quality of the best response by category scoring, (4) number of responses, (5) number of unusual responses, and (6) number of responses that are both unusual and of high quality. The tests were administered to about 4,000 applicants for admission to graduate school, using an item-sampling procedure. The tests were found to be appropriate in difficulty for advanced students. Reliabilities of most scores were high enough to be useful. Factoring of score intercorrelations reveals a general number-of-responses factor and two quality factors that are defined by quality scores from different combinations of tests. The number scores are quite independent of conventional aptitude and achievement tests, and quality scores have a substantial amount of true variance not predicted by aptitude and achievement tests. The face validity of the tests seems to appeal to students and teachers, but evidence of construct validity is needed.

Table of Contents

Page

Plan of the Study 3

Rationale for Test Development 4

Description of the Tests of Scientific Thinking 8

Scores and Scoring Methods 14

The GRE Administration Study 22
    Description of Sample 24
    Scoring Reliability 26
    Test Reliability 28
    Means and Standard Deviations 30
    Correlations of Scores Within Each Test 34
    Correlations of Scores from Different Tests 38
    Correlations with GRE Scores 41
    Factor Analysis 44
    Smallest Space Analysis 49
    Relations with Background Information Questions 52
    Reanalysis for "Select" Sample 56
    Reanalysis for Effects of Item Order 62

The University of Washington Study 64

Medical School Study 70

Summary 73

References 81

Footnotes 83

Tables 1-28 85

Figure 1 115

Appendix A: Directions and Sample Items 116

Appendix B: Instructions for Scoring the Tests of Scientific Thinking 125

Appendix C: Statistical Procedures 139

DEVELOPMENT OF MEASURES FOR THE STUDY OF CREATIVITY[1]

There is no question that there are occasional individuals who stand out because of an unusual ability to suggest new and unusual solutions to problems, to invent superior methods for accomplishing tasks, to ask questions that put problems in a striking new perspective--individuals who are creative. Creativity is a quality sought after in selecting graduate students and recruiting teaching and research personnel, whether for work in natural sciences, behavioral sciences, or the humanities. If we better understood the phenomenon of creativity we might not only do a better job in identifying individuals who are likely to display the qualities we desire in their research, writing, or artistic production, but we might also learn to facilitate the development of creativity in homes and schools.

The scientific investigation of creative behavior would be greatly facilitated if we possessed a convenient and effective way of measuring that quality--if we had a standard set of situations that would elicit responses that can be characterized as creative or not creative. In other words, development of good criterion measures would be a desirable first step in studies of the creative process.

The purpose of this project is to develop a set of tests that may serve, at least provisionally, as criterion measures in subsequent investigations of creativity. We begin with attempts to measure creativity in the area of behavioral science (more specifically, psychology), with the idea that, if successful, the methods may later be extended to other sciences and possibly to other areas such as artistic production and creative writing. The scope of the present project is (1) to develop a set of tests to elicit creative performance and (2) to assess the acceptability of the tests from the standpoint of their psychometric properties (reliability, difficulty, intercorrelations) and their construct and discriminant validity.

The availability of a set of measures that successfully assess various aspects of creative performance would make possible a whole range of future investigations dealing with the processes involved in being creative. Some of these studies would be correlational, others experimental, and some would involve both correlational and experimental methods. In the correlational studies, one could investigate relationships of various aspects of creativity to many individual characteristics (e.g., cognitive abilities, motivation, personality, cognitive styles, and biographical information). The experimental investigations might deal with incentives, stress, training methods, climates of research organizations, opportunity for incubation of ideas, presence of models to emulate (in the form of creative researchers and teachers), and so on. Such a program of research should lead eventually to the development of a theoretical model of the creative process--a model that should prove useful in improving the selection and training of students, in maximizing the productivity of research scholars, and in contributing to a scientific understanding of human performance.

Plan of the Study

The development and evaluation of instruments whose scores would serve as dependent measures in studies of the creative process involves two parts: (1) the development, tryout, and evaluation of a set of tests primarily from the standpoint of their psychometric properties--their reliability, intercorrelations, speededness, item difficulty, etc.; and (2) exploration of the relationships of these tests to other variables that throw light on construct and discriminant validity.

It was originally planned to administer the experimental tests, along with other measures of ability, personality, cognitive style, and the like, to students in about their second year of graduate study. Such arrangements were in fact made at one university, and the results of that substudy will be described later. It soon became evident, however, that finding large enough groups of graduate students who were willing to provide the necessary testing time was difficult, and that the first step could be carried out more expeditiously if a larger group of examinees could be found. The solution was to make use of the time reserved for pretesting new items in the GRE Advanced Psychology Test. The procedure used made it possible to try out four 6-item tests, using an item sampling method, and it provided several additional measures useful in evaluating construct and discriminant validity. These included the GRE aptitude tests, the Advanced Psychology

Test with its subtest scores, and responses to questionnaire items dealing with amount and kind of training attained at the time of testing and the amount and kind of further training planned. Thus through one larger study it became possible to combine the study of psychometric properties with a limited investigation of construct and discriminant validity. Moreover, the sample for this study constituted a cohort of students planning to embark on a career in psychology. The use of such a group introduces the possibility of follow-up studies of the predictive validity of the tests for graduate school and later performance.

Rationale for Test Development

The two major approaches to measuring creativity that have been employed are those typified by the work of Guilford and MacKinnon. Guilford (1967) has developed a theory of cognition that he calls the "structure of intellect." This structure is represented by a solid figure, the three dimensions of which represent four kinds of *content* (figural, symbolic, semantic, and behavioral); five psychological *operations* (cognition, memory, divergent production, convergent production, and evaluation); and six *products* of information (units, classes, relations, systems, transformations, and implications). The model thus implies the existence of 4 x 5 x 6 or 120 unique abilities, such as the "divergent production of semantic units." This particular ability is often called *flexibility*; it is measured, for example, by tests called *Consequences* ("What would be the consequences if people no longer wanted or needed sleep?") and Plot Titles (list possible titles for a story plot that is presented to subjects). These tests are used by some investigators as criterion measures in studies of creativity. This general approach is an extremely analytic one that is useful in providing hypotheses about the nature of creativity, but the tests themselves can hardly be used as criteria of creative performance without begging the question. Development of criteria of scientific creativity requires that we use measurement devices that resemble more closely the situations in which creative performance is likely to occur in real life.

The other approach, represented by the work of MacKinnon (1962) and Barron (1965), is to choose people who are recognized as creative contributors in a particular field, such as mathematics or architecture. This is done by having members of the appropriate professional group nominate their most creative members. These individuals are then invited to spend several days in a period of intensive assessment. A similar assessment is conducted for members of a less outstanding group of practitioners of the same profession. The differences in assessments are used to define the characteristics of creative individuals.

Neither approach is satisfactory from the standpoint of our present objective of developing provisional criterion measures. The first is too analytic; it provides ideas and tests that are potentially useful in later correlational studies, but few people would be willing to accept Consequences or Plot Titles for the purposes we envision. The assessment procedures, on the other hand, are not feasible for the kind of investigations we propose, although the comparison of highly creative with more ordinary people would be useful in helping to validate our provisional measures.

Real-life data are awkward to work with scientifically because of the lack of control; variations in performance may be attributed to any of a host of variations in opportunity and occupational situations in which people are employed. It is precisely for such reasons that many psychologists retreat to the laboratory, where they can control rigorously many of the factors that influence the dependent variables in their studies.

Our strategy involves a middle ground between the extremes of real-life criteria and the analytic procedures that may provide hypotheses about correlates rather than measures of creative performance. We wish to develop measures that resemble quite closely the real-life problems of a scientist, while retaining some of the control of a laboratory experiment. This we propose to do by using tests of the type that are sometimes referred to as work-sample tests or situational tests. For our purposes the most useful work in this area is that of John Flanagan (1949), who studied critical requirements for research personnel employed in 20 research laboratories. Of Flanagan's list of eight critical requirements, three seem particularly relevant: Formulating problems and hypotheses; Planning and designing the investigation; and Interpreting research results.

Let us imagine a scientist who is inspecting some charts and tables depicting an unexpected result of an experiment he has conducted. He will probably begin by formulating a clear statement of the finding. He may then review the study, looking for sources of error such as confounding of variables, errors in research design, or inappropriate method of analysis, that might have produced the finding. If he finds no likely source of error he may next engage in speculations about new theoretical implications and insights suggested by the data. Eventually he will have to recognize the constraint that whatever interpretation he seriously entertains should be consistent with all the data and with other information available to him. After engaging for a while in the process of generating explanatory concepts, on the one hand, and evaluating them against such criteria as theoretical and logical consistency, on the other, the scientist may be left with a number of hypotheses in mind that vary considerably with regard to the probability of their being correct. If at this point he is asked to summarize the implications of his study, he may, depending upon his personal predilections, propose a large number of interpretations, some of which are highly speculative, or he may propose only a few ideas that are "safe" in that they meet conventional standards of rigor.

Such an armchair analysis of a scientist's thinking suggests that several cognitive abilities may be involved. One kind of ability would presumably be ideational fluency, the ability to generate many ideas, including some unusual or original ideas. But because of the constraint that the solutions must be consistent with other known facts, other cognitive abilities such as reasoning and memory, as well as relevant information, will be required. This implies that a balance must be maintained between generating many ideas (including some that may be inconsequential) and discarding ideas (some of which may be valuable) that fail to meet the scientist's standard with regard to rigor or probability of veridicality.

Examining individual differences that may be manifested in arriving at such a balance gets us into the domain of personality. One way to account for such differences may be in terms of self-confidence or self-esteem. Those who are willing to propose interesting though implausible hypotheses may be people with a high degree of self-confidence and a lack of concern about the opinions of others. Those who discard all but the most obvious hypotheses may be anxious or defensive. Another way may involve differences in the kinds and levels of standards of excellence that one has learned to set for himself--standards that determine for each individual when to say, "That's good enough," "That's too risky," "That will attract attention!" or "That's a long shot, but it's worth trying."

Conceptions of this sort fall far short of constituting a theory of creativity, but such notions have helped to guide the development of tests so far.

Description of the Tests of Scientific Thinking

Prototype items of various kinds were developed and pretested informally. The four types which were eventually chosen for use in the GRE administration investigation may be described as follows:

1. Formulating Hypotheses (FH). Each problem consists of a brief description of a research study, a graph or table showing the principal results, and a statement of the major finding. The task is to write hypotheses that might explain, or help to explain, the finding. The individual is asked to write not only the hypothesis he thinks is most

likely to be correct but also other hypotheses that ought to be considered in interpreting the data or in planning another investigation. He is asked to mark the hypothesis he thinks is the best one.

2. Evaluating Proposals (EP). The examinee is asked to suppose that he is teaching a senior course in design and methodology, and as a class exercise he has asked his students to write brief proposals of research studies. Several of these proposals are presented as the test items. The task is to write critical comments to each student regarding the design, methodology, or theoretical position taken.

3. Solving Methodological Problems (SMP). Each problem is a brief statement of a methodological problem encountered by a graduate student or a psychologist in planning a research study. The task is to write suggested solutions to the methodological problem.

4. Measuring Constructs (MC). Each problem consists of a name and definition of a psychological construct (e.g., conservatism, bigotry, leadership ability). The task is to suggest methods for eliciting the behavior so that it can be observed and measured, without resorting to self-report methods or ratings by others.

For each test, the directions are followed by a sample problem and a sample answer sheet filled out by a hypothetical student. The sample responses have been carefully chosen to suggest the kinds of thinking the particular test is intended to elicit. (See Appendix A.)

Other tests which were considered include:

1. Analyzing Constructs. The directions for this test point out that certain psychological constructs that at first seem unitary turn out

on closer examination to break down into a number of separate (although possibly correlated) components. General intelligence is used as an example; intelligence has been shown to be composed of a number of separate abilities such as verbal ability, inductive reasoning, spatial ability, and so on. The student is asked in this test to name the components that he thinks might be found by appropriate research procedures for various constructs, such as socio-economic status and curiosity.

2. Evaluating Manuscripts. Each problem consists of a short manuscript that has supposedly been submitted to a journal for publication. The student is asked to assume that he has been asked to serve as a referee for the article, and is asked to write his recommendation regarding publication, statements of his major criticisms, and his recommendations to the author for revision (much in the style typically used in soliciting comments from a referee). Manuscripts are edited to be brief and nontechnical and to elicit various kinds of criticisms and comments.

3. Formulating Research Ideas. In this test, the student is asked to suppose that he is a graduate student who is trying to decide between two areas of specialization and that his advisor has suggested that, in order to get a better impression of the nature and variety of research projects he might engage in, he write down as many ideas for a dissertation project as he can think of in each field.

4. Personnel Selection Problems. Here the student is asked to suppose that he has accepted a job as research assistant for a firm that specializes in research on personnel problems in industry. The firm has just signed a contract with a large nonunion contractor to develop methods

for hiring employees in the building trades. The psychologist in charge of the project asks his new assistant to make a list of all the personal characteristics he can think of that might be associated with success as, say, a plumber. Other problems are similar, but with different jobs involved.

5. Ideational Fluency in Psychology. This test is very similar to a number of tests employed by Guilford and others to measure ideational fluency. The only difference is that psychological concepts are involved. The task is to "write as many words or phrases as you can think of that have been used to describe ____________." The blank can be filled by a term such as emotions, learning, or personality.

6. Evaluating Hypotheses. This is the only test that could be scored by machine. The same problems that are employed in Formulating Hypotheses are used, but in this test a list of five hypotheses are presented as multiple-choice options. The five hypotheses are chosen because they vary systematically in quality, using the ratings that are the basis for scoring Formulating Hypotheses. Alternatively, the student could be asked to rank the listed hypotheses in order of their likelihood of being correct. We have no plans at present for using such machine-scorable tests because there is little likelihood that any aspect of originality or ideational fluency can be elicted. But many of the tests could be adapted to multiple-choice form for other testing purposes.

It will be noted that the 10 tests described above cover a broad spectrum with regard to the degree of constraint imposed on the student in responding to the problem. At one extreme are tests (best exemplified by Ideational Fluency in Psychology) where constraints are minimal; in the example, the stimulus terms are taken from psychology, but there is no requirement that responses be specifically technical or psychological. Thus the test differs very little from the Guilford (1964, 1967) tests of word fluency or "divergent production of semantic units." (For this reason, we do not plan to use this test in developing "criterion" measures, although it may be useful in other ways.) Another test that appears to be near the fluency end of the spectrum is Personnel Selection Problems. Formulating Hypotheses has a fairly large component of fluency, as we know from studies with earlier forms of the tests, but the constraints are of considerable importance.

At the other end of the continuum would no doubt be Evaluating Hypotheses; since there is no opportunity for original responses, the test cannot measure fluency at all. Of the free-response tests, Solving Methodological Problems perhaps imposes the most constraint, although the student must still originate the ideas he writes down. Evaluating Proposals might also be relatively close to the "constraint" end of the spectrum, where presumably reasoning and knowledge would have greater importance than fluency.

The first four tests--Formulating Hypotheses, Evaluating Proposals, Solving Methodological Problems, and Measuring Constructs--were chosen for use in the present investigation because they appeared to represent major aspects of the job of a scientist, as revealed by the Flanagan critical incidents study, while varying widely along the continuum of degree of constraint imposed.

Responses to these job-sample tests obviously are complex. They will reflect originality and flexibility, if we are at all successful, but they will also be influenced by other cognitive abilities, as in rejecting ideas on the basis of knowledge and reasoning. They may also reflect temperamental characteristics that might, for example, result in self-censorship of ideas. It is this complexity (approaching, we hope, the complexity of real-life performance) that will make the responses useful for the later studies of creative processes that are contemplated.

To make effective use of these complex responses, it will be necessary to develop methods of scoring which are sensitive to several aspects of an individual's performance. Methods tried out with earlier versions of Formulating Hypotheses (designed for the general undergraduate population rather than with specifically psychological content) show that this is feasible. Scores intended to measure respectively the quantity and the quality of responses were both reasonably reliable, while the correlation between them was low and slightly negative. These results show that at least two independent aspects of productive thinking can be identified in a given set of protocols.

Scores and Scoring Methods

Two methods of scoring for mean response quality were tried in a recent study (Frederiksen & Evans, 1974). One method was to ask the scorer to make subjective evaluations, using a 9-point rating scale, of the quality of each response. The second method involved providing the scorer with a list of categories representing the ideas commonly produced by examinees, based on a classification of the responses in a sample of protocols, and asking the scorer to assign each response to the category it most resembled. The categories on the list were rated for quality by each member of a panel of judges, and a scale value based on these evaluations was assigned to each idea on the list. Thus it was possible to have a computer assign the appropriate scale value to each response given by a candidate. (If a response fit none of the categories, the scorer rated its quality on a 9-point scale. These ratings were later rescaled to be consistent with the distribution of scale values for that item.)

It was found in the earlier study that the quality scores obtained by the two methods measured essentially the same thing: the correlation between the two scores, for an N of almost 400, was .98 when corrected for the unreliability of the two scores. On the basis of this finding, it was decided to use only the latter method of scoring for quality.

This method appears to have several advantages. First, it requires less exercise of judgment on the part of the scorer; it is therefore both faster and less subject to differences in scorers' interpretations of the problem than the rating method. Second, it is less likely to be influenced by the length, neatness, or grammatical correctness of a particular response; the scorer's attention is focused on the meaning of the response rather than on irrelevant aspects of its form. Finally, it makes possible a range of modifications and extensions of the derived scores without requiring reexamination of the raw protocols. Any desired change in the quality scores can be obtained simply by entering into the computer a new set of values for the categories associated with each item.

The scoring method that has evolved gives rise to six scores for each test item. Three of these are directly concerned with the quality of the responses which are given to the item; two represent counts of numbers of responses; and one involves a combination of these two major aspects of performance. The six scores are as follows:

1. Mean Quality: The average of the quality values assigned to each response to an item.
2. Highest Quality: The quality value of the response with the highest scale value according to our scoring system.
3. Best Quality: The quality value of the best response to an item, according to the examinee's assessment of his own performance. (The examinee is asked, after completing an item, to mark the one of his responses which he thinks is his best. This score has been obtained for items from each test except Evaluating Proposals, where it was felt that the miscellany of bases on which a proposal could be criticized made such a comparison questionable.)
4. Number of Responses: The total number of scorable, non-duplicate ideas given as responses to an item.

5. Number of Unusual Responses: The number of responses to an item which meet a criterion of infrequency in the distribution of all responses to that item. Generally, an unusual response is one in a category which accounts for no more than 5% of all responses given by all examinees.

6. Number of Unusual High Quality Responses: The number of responses to an item which meet both the criterion for inclusion in "Number of Unusual Responses" and a quality criterion. The quality criterion is that the category into which the response falls is rated in the top third of all categories for the item.

These scores by no means constitute six independent dimensions of task performance; subsets of them have both logical and experimental interdependencies. As a set, however, they appear to provide a fair representation of the ways in which the responses given by one examinee differ from those given by another.

The Mean Quality score is the most straightforward representation of the individual's competence in dealing with the problem posed in an item. Highest Quality is of interest since this score could provide an index of the best performance of which the individual is capable. If, for example, he produces only one very good response to a problem followed by several which miss the point, his Mean Quality score will be lowered in direct proportion to the number of such additional ideas he writes; but the value of the Highest Quality response will be unaffected. This score, in other words, provides maximum credit to the individual for his occasional excellence without being affected by the volume of more pedestrian attempts he makes.

Best Quality, the quality of the response which the examinee considers his best idea, emphasizes the examinee's critical and evaluative abilities in addition to his productive thinking abilities.

Number of Responses, on the other hand, gives almost no weight to the quality of the idea, although a response must meet a minimum criterion of being comprehensible and non-duplicative of other responses given by the examinee. A history of high productivity seems to be characteristic of individuals who produce excellent products. This score is therefore of interest in its own right as an additional characteristic of productive thinking. In addition, it will provide a sort of control for the sheer fluency of response, a variable whose effects may confound some of the other scores from the test.

Number of Unusual Responses and Number of Unusual High Quality Responses are each subsets of the total number of responses, and may represent scores that are too confounded with the latter, or which involve too rare events, for high score reliability or validity. Each, however, is of interest in providing directly a way of looking at task performance which matches a frequently-found operationalization for "creative" production.

The category-based scoring for mean quality described earlier allows for computer generation of these six scores, once the list of categories for an examinee's responses to an item has been recorded. It will also allow construction of subscores for specific kinds of problems (e.g., for items based on controlled experiments versus those based on field studies) and for specific kinds of responses (e.g., methodological or theoretical responses).

While the category-based scoring system has many advantages, its usefulness presupposes careful development of the list of categories to be used. The speed and accuracy of a scorer depend upon the existence of a set of statements which are both comprehensive, in including all or nearly all responses likely to be encountered for an item, and nonoverlapping and unambiguous, so that there is little uncertainty as to which category a given response should be assigned.

There are two aspects to the development of response categories. The first is to develop a classification of the responses of a typical group of examinees. The second phase involves writing generalized statements that (ideally) are broad enough to include any response that falls in a particular category but which exclude responses that belong in other categories.

Development of the set of categories for each problem involved the coordinated efforts of both principal investigators and two experienced research assistants. The following sequence of steps was taken: (1) Two of the four individuals, working independently, took 50 protocols for one problem from the GRE data and constructed a trial set of categories which was intended to subsume all responses given by candidates in that set, along with any other responses the investigator thought might occur. (2) The two compared their lists and arrived at a consensus list. (3) The remaining two individuals, working independently, attempted to assign each response from the 50 protocols to a category on the consensus list. (4) Examination of disagreements, ambiguities, and failures to find a category into which a response would fit led to further changes, until all four individuals reached consensus on the category list. (5) When all six problems representing one of the tests had been processed in this way, one of the assistants reviewed the lists and suggested minor changes to achieve consistency of style and language across all the problems in the test. (6) Two scorers were given the list of categories along with a new set of protocols from 50 individuals who had taken the problem in the GRE administration. Their categorizations were compared by an assistant, and the disagreements and problems in interpretation led to further additions and clarifications. This sequence of activities was time consuming and occasionally exasperating, but led to a set of categories that have proved able to accommodate approximately 95% of all responses obtained in the GRE testing, with relatively few scoring problems.

To derive scores from the categorized protocols, a table of scale values representing the quality score to be assigned to each category is required. Again both investigators and two assistants have participated in the scale development. Each of the four individuals independently ranked the categories for each problem in order of quality, which was defined in terms of the instructions given examinees for the particular test. For the Formulating Hypotheses items, for example, the highest ranks were given for ideas judged most likely to provide a correct explanation for the finding, or to deserve serious consideration as competing hypotheses. Lower ranks were given for ideas seen as less plausibly true, or less likely to be the major contributor to the outcome even if true. The lowest ranks were reserved for those which were judged quite likely to be false or to be irrelevant to the outcome. In all cases rankings depended far more on judgment, conditioned by general knowledge of psychological principles, than on knowledge of specifically relevant results in the psychological literature; none of the problems were such that a single "pat" solution was derivable from previous work in the domain.

Agreement among the four judges[2] was in general very high, as is indicated by the alpha coefficients shown in Table 1. The coefficients ranged from .76 to .98, with a median value of .92.

Insert Table 1 about here

The lowest coefficients in Table 1 are those for Solving Methdological Problems Item 1 and Measuring Constructs Items 1 and 4. Measuring Constructs Item 2 is also relatively low. These four items elicited responses that seemed to require a somewhat different method of categorizing responses than was typically employed. Each Measuring Constructs problem presented the name and definition of a psychological construct, and the task was to suggest ways of eliciting the relevant behavior so that it could be observed and measured. The responses very frequently had two parts: (1) a proposed situation to elicit the behavior and (2) a proposed method for measuring or evaluating the behavior that was elicited. Lists of situations and lists of measurement methods were developed separately, and the scorer was asked to classify the response, whenever appropriate, along both dimensions. This procedure resulted in a much larger number of categories, since most of the proposed measurement methods could reasonably be applied to most of the proposed situations.

Because the number of categories for these items was large, they were rated on a 21-point scale rather than ranked. It is not clear exactly why the different procedure resulted in lower alpha coefficients but presumably it was related to the much larger number of categories that had to be kept in mind in making the judgments, as well as whatever differences are involved in the process of ranking as compared with rating. Overall, the amount of agreement in evaluating response categories was quite satisfactory.

The GRE Administration Study

With the approval of the GRE Advanced Psychology Test Committee, arrangements were made to administer four 6-item tests (Formulating Hypotheses, Evaluating Proposals, Solving Methodological Problems, and Measuring Constructs) to all candidates taking the psychology test in the United States in October of 1973.[3] The testing time reserved for pretesting new items--25 minutes--was used. The inclusion of all of our 24 items was made possible by using an item-sampling procedure.

The rationale for item sampling requires that subsets of items be administered to subgroups of candidates. If these subgroups are randomly chosen, variances and covariances of items obtained for each subgroup provide estimates of those that would have been obtained from the total group. Thus item variance-covariance matrices may be assembled and treated as though all candidates had taken all items. From these matrices unbiased estimates may be made of the reliabilities and intercorrelations of the 6-item tests, and by adding to the matrix the variances and covariances involving the various GRE scores the correlations of the four tests with GRE scores may also be estimated.[4] Relationships involving questionnaire items may be investigated by using analysis of variance procedures.

It was planned that about 40 percent of the candidates (the reliability sample) would be given three items from a single test, and about 60 percent (the intercorrelation sample) would be given two items from two different tests. (Taking two items required reading two sets of instructions instead of one, and studying two sample items.) Our formal and informal tryouts of the tests showed that seven or eight minutes per item was thought to be sufficient by most students; therefore the time allowance was believed to be adequate. Items were not separately timed, but students were informed that they had 25 minutes to read the instructions, study the sample item (or items) and write their answers. It was assumed that time per item would be approximately the same for the reliability and intercorrelation samples.

Two-item tests and three-item tests were assembled in such a way that all the combinations of two items from different tests, in all the possible orders, were present in equal numbers, as were all those of three items from one test. These experimental tests were then arranged in random order and placed in the GRE Advanced Psychology Test booklets.

An assumption involved in this procedure is that performance on a test item will not be influenced in any systematic way by the context in which it is presented--whether the item is presented with another item from a different test or with two other items from the same test. The assumption can be tested by comparing means and standard deviations of items administered under the two conditions.

Candidates were told in the instructions for Section V that "the time usually allotted to the tryout of new multiple-choice items has been consolidated, so that Section V can be used to try out items of a different type. Information from this tryout will be used in planning revisions of the GRE test and in studies relating to graduate education.

Scores obtained from Section V will not be used in obtaining your GRE Advanced Psychology Test score and will not be reported to anyone. The identifying information will be used only to make it possible to do correlational studies involving other information obtained through this administration of the test. Your signature on the line below will indicate your permission for these uses of the information you provide by taking Section V." Thus a full disclosure was made to the student. We also asked his permission for us to request his participation in a later follow-up investigation, if one should be carried out.

The October 1973 testing went according to plan. Supervisors at the testing centers generally reported no difficulty. We had had some concern about the extent of cooperation, in view of the disclosure that the experimental tests would not affect test scores, but except for two centers where there was widespread nonparticipation most students not only took the tests but also indicated willingness to participate in a follow-up study.

Description of Sample

The total number of students given the Advanced Psychology Test in domestic test centers was 4,394. The number of candidates with complete data who were actually used in the analyses was 3,586. Table 2 lists the reasons for loss of the remaining 808 examinees; approximately half of these were refusals, while the others represent cooperative individuals whose data were incomplete or spoiled.

Insert Table 2 about here

All candidates taking the GRE Advanced Psychology Test are routinely asked to respond to nine Background Information Questions which deal with the amount and kind of training the individual has had and with his intentions for further training. The distributions of answers to these questions will be useful both in characterizing the sample and in showing to what extent it differed from the entire October 1973 administration group. As seen in Table 3, the typical candidate was about what one would expect: a senior psychology major planning to attain a doctorate in psychology. More than two-thirds of the group had training in general and experimental psychology and in statistics. Just over half planned a career in clinical psychology. Those in our complete data sample differed hardly at all from the total group in these respects. To the extent that there were differences, the sample was slightly more like the modal GRE candidate: one to three percent more were undergraduate seniors, were majoring in psychology, and were planning for a doctorate in psychology.

Insert Table 3 about here

It is also possible to compare the entire October group with the sample on the basis of scores on the GRE aptitude and achievement tests. The means and standard deviations are shown in Table 4. For all tests and subtests the means are slightly higher for the sample. Apparently the students who chose not to take the experimental tests tended to be those who were slightly less conventional in their educational careers and who were slightly less able in test performance. However, all differences are so small that we can safely generalize to the candidate population.

Insert Table 4 about here

The GRE scores may also be used in determining the comparability of the major subgroups within the complete data sample. The item-sampling plan was intended to produce 10 equivalent groups. Four of these groups, constituting the reliability sample, are the groups containing all candidates who were given three items from one of the four tests (FH, EP, SMP, and MC). The remaining six groups, collectively making up the intercorrelation sample, are composed of all candidates who were given two items, one from each of two of the tests; each group represents those receiving items from one of the six possible test pairs.

One-way analyses of variance were conducted to determine whether the random assignment of candidates to groups did indeed produce equivalent groups. Using the aptitude scores as dependent variables, with 9 and 2,987 degrees of freedom, *F* is 0.38 for the Verbal score and 1.17 for the Quantitative score. When the Advanced Psychology subscores and total score serve as dependent variables, *F*'s, with 9 and 3,550 degrees of freedom, range from 0.68 to 0.87. Since an *F* of 1.88 is required for significance at the .05 level, the groups do not appear to differ appreciably from one another.

Scoring Reliability

Scoring was done by part-time at-home workers, all of whom had bachelor's or master's level training in psychology or a closely related discipline (e.g., educational psychology, sociology). Two scorers were assigned to each item, and each completed all the protocols for that item before being trained to score another one. (Instructions to scorers and a sample score sheet are included in Appendix B.) An assistant checked each set of 50 protocols when it was returned, to allow continuing monitoring of the scoring process. After scoring had been completed, the categorized protocols were keypunched and stored on magnetic tape. A computer program was written which derives all of the six scores for each of the 24 items, separately for each scorer and for the two scorers combined.

The scoring reliability (coefficient alpha) for each item is shown in Table 5. These coefficients represent the reliabilities of scores based on the judgments of two independent scorers.

Insert Table 5 about here

The median item reliability (over all four tests) for number of responses is .90, which, as might be expected, is the highest of the scorer reliabilities. It might be wondered why the reliability of counting responses should not be still higher, but it must be remembered that judgment is involved even here, for example in deciding whether a given answer actually contains two different ideas.

It is interesting to note that scorer reliability for Best response is almost as high as for Mean quality of responses and that for Highest quality it is actually greater than for Mean quality. This is true even though the Mean score is based on all the answers while the other quality scores are derived from a single response for each item. One possible explanation is that the best answers tend to be more clearly formulated than the responses of lower quality and hence easier for scorers to classify correctly. Another possibility is that, since categories with high scale values often occur with high frequency, scorers have more experience in dealing with them.

Reliabilities for Unusual and Unusual-High responses are the lowest, as is to be expected for two reasons: (1) they are based on small subsets of responses, and (2) they are based on responses that correspond to categories that scorers have little experience with--or ideas that may not be found in the list of categories at all.

Scorer reliability is high enough, even for Unusual and Unusual-High responses, to justify the use of all of the scores in further explorations of the psychometric properties of the tests. If one uses the Spearman-Brown prophesy formula to predict the reliability of scoring the 6-item tests from the median item reliabilities, the reliability is found to be .87 even for Unusual-High.

Test Reliability

The reliabilities of the six scores for the four Tests of Scientific Thinking are shown in Table 6. These coefficients necessarily reflect both scorer agreement and amount of consistency in performance on the part of the examinees. They are estimated reliabilities for 6-item tests, calculated under the assumption that each subsample of individuals given a particular pair of items from one of the tests provides an unbiased estimate of the result which would have been obtained had all individuals been given a 6-item test. Here, and in the calculation of other estimated coefficients for 6-item tests, if individuals given the various subsets of items were perfectly comparable, the estimates derived would be exactly those which would have been yielded by complete 6-item tests.

Insert Table 6 about here

Two estimates are given in each cell of the table. The first entry is a lower bound reliability estimate. The second entry may be thought of as an upper bound coefficient; it is an estimated parallel-form test-retest reliability, assuming two forms that are maximally similar (a condition that could not actually occur). The methods for obtaining these reliabilities are described more fully in Appendix C.

The upper bound estimate of reliability will be useful in providing conservative estimates of true-score correlations of the experimental tests with each other and with other measures.

The lower bound estimate is the one that corresponds more closely to conventional methods of computing reliability; and since it is also the more conservative estimate from the standpoint of a psychometric evaluation of the tests, it will be used in summarizing the information on reliability. The differences between lower and upper bound estimates vary from .05 (for MC Number) to .21 (for SMP Best). The median difference is .13.

In comparing tests with regard to reliability, it is apparent that SMP is the lowest for all scores except for Unusual, and MC is highest except for Unusual and Unusual-High. It is more difficult to generalize about the reliabilities of the six scores, since reliabilities vary from test to test. The Number score tends to be most reliable and most consistent across tests, as might be expected for a score based on a count of the responses. The Unusual-High reliability is low for MC and very low for SMP, but it is high enough to be useful in the cases of FH and EP. The Unusual score (which is related to Unusual-High, in that the latter score is based on a subset of Unusual responses), is also relatively low in reliability for the SMP and MC.

Means and Standard Deviations

The means and standard deviations of the scores, based on the reliability sample, are shown in Table 7. The mean shown for each Number score is actually the average number of nonduplicate responses per item. It is apparent that candidates had more ideas (almost four per item) for Evaluating Proposals than for any of the other tests. FH and SMP resulted in about two and a half ideas per item, on the average. It had been our impression that the problems posed in SMP items were intrinsically the most difficult; apparently FH is a close second, judging from the number of ideas produced.

Insert Table 7 about here

On the average, the number of responses classified as Unusual was roughly a fourth of the total number of ideas per item, and the number of responses classified as Unusual-High was generally less than a third of that. The reliabilities of the count scores are related to the size of the mean scores, but by no means perfectly so.

The mean quality scores are not comparable across tests. One reason is that they are based on relative rather than absolute judgments of quality and are obtained from items that differ in difficulty in unknown ways. Another reason is that since the scale values are based on ranks and the highest possible rank has a value equal to the number of categories, the scale value of the best category is higher for the items with larger numbers of categories.

Table 8 shows the number of categories for each item in each of the four tests and also the theoretical upper limit for each of the three quality scores and for the Unusual and Unusual-High scores. These values may be helpful in understanding the quality of performance represented by the mean scores shown in Table 7.

Insert Table 8 about here

The distributions of Number of responses tend to be skewed toward the high end for most items, although the distributions were more nearly symmetrical for EP. The highest number of responses varied from a maximum of 5 for one item up to 11 for one EP item. For the Unusual score, the modal score was either 0 or 1; the highest score for any one item was 7, and for three items the largest number of Unusual responses was 2. For Unusual-High scores the modal score for every item was zero, and the highest score for any one item was 4.

Distributions of Mean Quality scores were either skewed toward the low end or were more or less symmetrical. For Best responses the score distributions were more often skewed toward the low end, although some were quite symmetrical. The skewness of Highest scores toward the low end was much more marked, as would be expected in the light of the method of scoring.

The score distributions described above are based on the reliability sample. The N's for single items ranged from 143 to 195 for all scores except Best. Some candidates neglected to designate the response they considered best for each item; the range of N's for this score was 108 to 176.

The means and standard deviations shown in Table 7 are based on the reliability sample, i.e., the candidates who took three items from one test. The intercorrelation sample consists of all those candidates who took two items, both from different tests. It was assumed, in setting up the procedure, that the time that could be devoted to an item would be approximately the same for both groups, since those who took only two items had to study two sets of instructions and two sample items instead of one, and therefore the means and variances would be about the same. Table 9 shows the means for the intercorrelation sample.

Insert Table 9 about here

A comparison with Table 7 reveals that for the three quality scores the differences in means are indeed small. The differences are systematic, in that the Best and the Mean Quality scores are lower for each test for the intercorrelation sample, while the Highest Quality scores all have higher means in this sample. However, in no case do the means differ by more than 4% (0.3 standard deviations). For the three scores involving counts of items, the means are all higher for the intercorrelation sample, and here the differences are sometimes quite large. Over the four tests, students taking only two items wrote on the average 15% more responses per item, gave 16% more unusual answers, and gave 15% more which were both unusual and of high quality.

These differences suggest that the time required to read the instructions and study the sample item for the second test was short enough to allow somewhat more time for responding to the items. This extra time resulted in the production of more responses without appreciably influencing the quality of the best ideas or even reducing the average quality. Apparently the best ideas tended not to be produced near the end of the allotted time, and the extra time resulted in more answers of only average quality.

The standard deviations reported in Table 9 require some comment. Estimation of the variance of a 6-item test cannot be based on the performance of the intercorrelation sample alone, since the estimate requires the use of covariance terms for pairs of items within a test. Those terms are, however, to be found in the data for the reliability sample. The study design therefore called for the use of estimated test variances from the latter part of the total sample wherever these were required in dealing with data from the intercorrelation sample.

For the quality scores, the comparability of overall means from the two groups, as well as the similarity of variances in scores at the item level (not presented), suggests that this procedure is justified; therefore, the standard deviations reported for quality scores in Table 9 are the same as those estimated from the performance of the reliability sample. For the count scores, both the means reported in Tables 7 and 9 and a comparison of item-level variances suggest that an item presented in a two-item context should be treated as a longer "test" than the same item given in a 3-item context. The ratio of means under the two sets of conditions provides an index of the amount of lengthening. By use of Gulliksen's (1950) formula for the variance of a lengthened test, this ratio provides a basis for the necessary correction in the test variances estimated in the reliability sample. The standard deviations for count scores in Table 9 are based on the corrected variances. Further detail is given in Appendix C. The importance of such corrections will be more apparent in later discussions.

Correlations of Scores Within Each Test

The correlations of the scores derived from any one of the tests can be computed from a matrix of variances and covariances of the six scores for each of the six items. Data supplied by the reliability sample provide all the terms needed. The estimates will be exactly equal to the coefficients which would have been obtained had each individual taken a complete 6-item test, so long as two conditions are met: (1) that individuals randomly assigned to the various subsets of the pool of six items are exactly comparable and (2) that performance on a given item is unaffected by its being given as one of three items rather than as one of six.

The correlations thus estimated will be seriously inflated by the presence of experimental dependencies among the several scores derived from responses to a single item. This is not a flaw in the estimation procedure, but rather an indication of the logical and empirical relations among the scores. Another procedure has been developed in which correlations are estimated using only covariance terms which reflect non-interdependent scores, thus eliminating the inflation of the coefficients.

Both methods were used, and the intercorrelations of the six scores for each of the four tests are shown in Tables 10 and 11. Table 10 shows the intercorrelations of scores obtained using all the covariances. These correlations are good estimates of what would be obtained if entire tests were administered without item sampling. Table 11 includes correlations using only the off-diagonal covariances, thus excluding the covariances based on use of the same protocols to obtain the six scores. These correlations may be thought of as estimates of the intercorrelations that would be obtained by giving a different set of test items to yield each different score. The correlations in Table 11 are lower because they are free of the experimental dependence, and they are therefore the ones that should be used in judging the degree of relationship among the abilities represented by the scores, in contrast to the scores themselves. The differences, generally speaking, are greater for the intercorrelations among quality scores and for the intercorrelations among count scores than they are for the correlations between quality and count scores.

Insert Tables 10 and 11 about here

Looking at Table 11, we see that the three quality scores tend to form a cluster (although for SMP the tendency is less marked). The three count scores also form a cluster in the cases of FH and EP. These clusters occur in spite of the fact that experimental dependencies have been eliminated. The Unusual-High scores fail to share substantial variance with the other two count scores for SMP and MC; quite possibly, this departure from the pattern of the remainder of the data is simply a reflection of the very low reliability of the Unusual-High scores for these two instruments. The correlations between the quality scores and the count scores tend to be lower, although there is no very consistent pattern; these correlations are all negative for SMP. One might guess that for SMP those who wrote responses of high quality (responses that really solved the problem) felt that there was not much point to adding more answers. Perhaps the items in the other tests were of such a nature that it was not as clear to the candidate when he had provided an adequate solution.

In Table 12 are shown estimates of what the correlations in Table 11 would be if the variables were perfectly reliable. They are based on "corrections for attenuation" using the "parallel forms" reliability coefficients, which provide the more conservative estimates of the true-score intercorrelations.

Insert Table 12 about here

An interpretation of the intercorrelations shown in Table 12 would not be different from the one we have given for Table 11. The clusters of related scores are more obvious in Table 12, but in general the pattern of interrelationships of the scores is the same. The magnitude of the correlations does not indicate that there is a great deal of redundancy in the information provided by the six scores. The greatest amount of redundancy is in the quality cluster, and even here no single score accounts for more than about half of the variance in another score.

Correlations of Scores from Different Tests

Sixty percent of the candidates tested made up the intercorrelation sample, intended to provide estimates of the correlations among scores from the four Tests of Scientific Thinking. One-sixth of this group was assigned to each of the possible pairs of tests. Within each such subgroup, each candidate was given one item from each of the two tests. All possible item pairs were administered in both possible orders.

Computation of the correlation between any two scores requires use of the interitem covariances and of estimated total test variances for those scores. The covariances present no difficulty; each of the 36 terms is based on a small number of cases (averaging nine to eleven when incomplete data cases are eliminated), but the sum of 36 such terms should provide a good estimate. Test variances, however, must be estimated using data from the reliability sample, as was discussed in an earlier section. The correlations presented here were derived using the procedure described in that section--that is, taking variance estimates for the quality scores directly from the reliability sample results, and for scores involving counts applying a correction for differences in test "length" to that sample's variances.

The entries in the 23 x 23 matrix (based on four tests, three of them providing six scores each and one providing five scores) that are of most interest are the correlations between corresponding scores from different tests. These correlations may indicate something about the complexity of scientific thinking--whether an individual's production of ideas or the quality of his ideas is uniform across different kinds of tasks, or whether the ability to produce solutions to problems is specific to the kind of problem posed. The estimated correlations are shown in Table 13.

Insert Table 13 about here

If the coefficients in the columns of Table 13 were uniformly high, it would be suggested that problem-solving performance is constant regardless of the type of task required of the subject, at least within the domain encompassed by the four tests. Moreover, it would indicate that we could combine items from the four test types and use just six scores on a composite "Test of Scientific Thinking," rather than four sets of six scores. In fact, the correlations vary quite widely, not only within each column but also within each row. It does not appear justifiable to combine similar scores obtained from different tests except in specific instances. Neither can one assume that a particular pair of tests will be comparable with regard to their ability to elicit similar behaviors over the whole spectrum of scores. In other words, there is a strong interaction between tests and scores. There is little support for a generalized ability to produce many ideas or good ideas.

One of the coefficients in Table 13 is obviously unreasonable, an estimated correlation between number of responses on FH and on SMP of 1.03. Table 14, giving true score correlation estimates derived using the entries in Table 13 and the parallel forms reliabilities of the scores, shows even more clearly that the assumptions underlying the estimation procedure were not entirely met; here three scattered coefficients are found which substantially exceed 1.0. Apparently test variances for scores in the intercorrelation sample have been underestimated in at least some cases.

Insert Table 14 about here

A second procedure was developed for estimating test variances for the intercorrelation sample. This procedure, described in Appendix C, was based on the assumption that the reliability of each score was identical for the intercorrelation and reliability samples; no other information from the latter sample was used. When variances derived in this manner were used in computing correlations, the resulting coefficients did not differ appreciably from those in Table 13. For example, the median correlation in Table 13 is .35, while the median for the comparable coefficients computed by the second procedure is .39. The two sets of coefficients over these 23 key test by score combinations themselves correlate .98. Thus, while the absolute magnitudes of the estimated correlations are suspect, the pattern of coefficients is quite stable over two methods which differ substantially in the assumptions on which they are based. Factor analyses of the correlations, to be reported in a later section, may thus help to clarify relations among these scores while removing spurious aspects of the magnitudes of the numbers.

Correlations with GRE Scores

Correlations with GRE aptitude and Advanced Psychology scores are of interest for the light they can shed on the construct and discriminant validity of the experimental tests. The new tests presumably require verbal comprehension and expression of ideas, reasoning from the terms of the problem given, and some understanding of facts and principles in psychology. Thus, they may be predicted to bear some relation to GRE Verbal, Quantitative, and Advanced Psychology scores. Moreover, differences in the magnitudes of their relations to these scores may provide an indication of the abilities or processes which are most involved in solving these problems in scientific thinking. On the other hand, to be useful the new tests must be discriminable from all the more conventional assessment indices; the time and effort required for test administration and scoring can only be justified if they provide additional information beyond that which can be more easily obtained from existing standardized tests.

Estimated correlations with GRE scores are presented in Table 15, and the corresponding true score correlations are given in Table 16. These coefficients were derived from the performance of the reliability sample only, since with this group test variances can be estimated without resort to the correction procedures which are required in dealing with intercorrelation sample data. _Ns_ for the analyses are somewhat smaller than those reported in Table 6. For the Advanced Psychology Test, only a few cases, those who requested cancellation of the test, are lost; for the aptitude scores, however, Ns are approximately 19% smaller, since a number of individuals did not take the Aptitude Test in the October 1973 administration. Most of these individuals have on file aptitude scores from other administrations, but these were not included in the present analyses. Reliabilities used in computing the true-score correlations are the upper bound estimates shown in Table 6 for the experimental tests, and the reliabilities for the GRE tests are those for the entire group of October 1973 candidates.

Insert Tables 15 and 16 about here

The first thing that strikes one in inspecting the tables is the uniformity among the correlations across the rows. The abilities represented in all the GRE tests are related nearly equally to all the scores and all the tests, although the contribution of verbal ability tends to be the greatest and that of knowledge of experimental psychology the least. Reasoning, as represented by the GRE Quantitative Aptitude Test, makes the second largest contribution, followed by the Advanced Psychology Test total score.

It is clear that the quality scores are in general more closely related to the GRE tests than are the count scores. This finding is consistent with earlier studies involving FH. However, reference to Table 16 shows that a substantial amount of true variance in quality scores is not accounted for by any GRE test; typically, a test accounts for about 25% of the variance (although the psychology test accounts for about two-thirds of the variance for highest quality score for SMP).

The scores based on counts do not share nearly as much variance with GRE tests as do the quality scores. Judging from previous studies, tests measuring abilities such as ideational fluency would be more related to Number and Unusual scores.

The GRE tests are not ideal tests for studying construct validity of the experimental measures because they by no means represent pure cognitive abilities. Both the verbal and quantitative test undoubtedly reflect reasoning ability, for example, in part because of efforts over the years to improve the predictive value of the tests. Similarly, the psychology test is not merely a test of knowledge; it contains items requiring applications and problem solving. It is perhaps for such reasons that their correlations with the new test are quite uniform across rows. But we can conclude that the new tests do not duplicate the existing tests with regard to abilities measured. There is a substantial amount of true variance that does not correspond to abilities presently measured by GRE scores even for quality measures, and the quantity and unusualness scores are still more independent of GRE measures.

SMP appears to be the test which most overlaps with GRE tests in what it measures, while MC perhaps shares the least variance with these instruments. SMP is also the least reliable of the experimental tests, and its scores have the most erratic relations with those of the other tests. In its present form, therefore, it is somewhat less useful than the other new instruments. However, the possibility must be considered that new items could be created which would better meet the requirements of construct and discriminant validity.

Factor Analysis

A series of factor analyses was undertaken in an attempt to clarify the structure of interrelationships among the various scores for the four tests.[5] The input was the 23 x 23 correlation matrix made up of six scores from each of three tests and five scores from the remaining test. Correlations among scores within each test were the coefficients from which experimental dependencies had been eliminated, and correlations among scores across tests were those computed using test variances corrected as described in the text above. The one estimated correlation in the matrix which exceeded 1.0 was replaced by the product of the square roots of the reliabilities of the two scores involved; this is equivalent to assuming that the true score correlation between the two variables is 1.0. Iterated communalities for some scores exceeded 1.0, which is another indication of inaccuracies attributable to the estimations required by item sampling. In order to avoid this problem, the square of the largest off-diagonal element in each column was used, without iteration, to estimate communalities.

Varimax (orthogonal) rotations based on three-, four-, five-, six-, and eight-factor solutions were examined. The three-factor solution is favored, both on the basis of the plot of magnitudes of successive roots and for reasons of interpretability. This solution is shown in Table 17.

Insert Table 17 about here

The first factor is clearly a general number of responses factor. The Number and Unusual scores from all four tests have their highest loadings on this factor (all $\geq .45$). Factor II appears to be a quality factor mainly for MC and FH, and Factor III is a quality factor based mainly on SMP and EP. There is one substantial loading on Factor II that does not conform to this pattern, that of Mean Quality on EP; and there are two loadings that do not conform on Factor III, a positive loading for Unusual-High on FH and a negative loading for Unusual on SMP. For each test, however, all the quality scores have their highest loadings on the same factor.

When more factors are retained, the factors shown in Table 17 tend to become more nearly test-specific. With five factors, for example, three quality factors are obtained--one primarily representing the three FH quality scores, one representing MC, and one combining EP and SMP. There are also two factors loaded only by the count scores. The first includes large loadings for the FH and EP count scores, and the second the MC count scores; but no pair of tests is consistently either separated or linked in the structure.

The four Unusual-High scores show no consistency in their placement in the factor structure, and two of them have extremely low communalities in these analyses. A series of analyses was run based on intercorrelation tables from which these four scores were deleted. Three-, four-, and five-factor varimax solutions were examined. The five GRE scores were included as extension variables, so that their relations to each factor could be observed without their influencing the factor structure.

Again the three-factor solution is favored. The results, which are presented in Table 18, match those of the analysis in Table 17 in all important respects. There is a single factor on which all the Number and Unsual scores have loadings above .40; and there are two quality factors, one which is primarily quality on FH and MC, and one which is primarily quality on EP and SMP.

Insert Table 18 about here

The loadings of GRE scores show that all tests have similar small relationships with the quantity factor. Correlations with Factor II are a bit higher, especially for the two aptitude tests; and they are still higher for Factor III, especially for the achievement test. Although the differences in correlations are probably too small to permit any firm generalizations, it would appear that EP and SMP require more knowledge of psychological facts and principles than do FH and MC, while the latter require more verbal ability and reasoning.

Three- and four-factor solutions were also obtained using oblimin, an oblique rotational procedure. The three-factor solution was highly similar to the varimax results. The two quality factors obtained by this method correlated .27, indicating only a slight degree of relation between them, while each was essentially independent of the number factor. The four-factor solution has no simple interpretation.

While none of these analyses is completely "clean," they do succeed in highlighting several aspects of the patterning of the scores. First, the several quality scores from a given test, even when statistically freed of experimental interdependencies, consistently appear together on the same factor. It is still possible that some method can be found to differentiate among quality scores; for example, the quality scores may differ in the degree to which they are related to an external criterion. However, it is clear that any such differences will be minor in relation to the overall coherence of these scores.

Second, response quality does not represent a single underlying ability dimension across the set of tests. There are at least two distinct abilities, and there is some slight indication that there are three. Quality of ideas on EP and SMP defines a single factor throughout the various analyses, and it contrasts with quality of ideas on FH and MC.

Third, in the count scores for each test, the Number and Unusual scores tend to cohere. The results here are reminiscent of those which have been obtained in studying performance on simple ideational fluency tasks (e.g., Ward, 1969), where the number of unusual responses appears to be a consequence of the rate at which the more obvious possibilities are exhausted, rather than representing a distinct process in itself.

Finally, the bases on which individuals are differentiated in idea quality are unrelated to those on which they may be distinguished in quantity of solutions. The Number and Unusual scores from a given test tend not to relate to the quality scores from that test; these count scores show a fair degree of coherence across all four tests, in contrast to the quality scores; and in those factor solutions in which the number factor does begin to separate, the division is not into the same test pairings as is obtained for the quality scores.

A major difficulty with traditional factor analysis is that the interpretation of results is largely a matter of art or intuition. An attempt was made to apply Jöreskog's maximum likelihood factor analysis to the correlation matrix (Jöreskog, Gruvaeus, & van Thillo, 1970). This method provides for a test of the significance of the fit of the obtained matrix to a hypothesized factor matrix. The maximum likelihood analysis, however, is extremely sensitive to inconsistencies in the matrix, and could not produce a solution. A method suggested by Tucker (personal communication) for resolving these inconsistencies was tried, in which a pseudo-complete data matrix was created by assigning every candidate the sample mean value on every score he was missing. This method has been useful with data sets containing up to 40% missing data; but, viewed in these terms, our item-sampling data set has 90% missing data, and satisfactory results were not obtained.

Smallest Space Analysis

Another approach to structural analysis is given by nonmetric procedures like the Smallest Space Analysis (Guttman, 1968). This analysis takes as input a matrix of coefficients that can be interpreted as representing distances between pairs of scores. It attempts to preserve ordinal relations among scores (rather than, as in factor analysis, preserving interval relations) in a space of reduced dimensionality. Since a less restrictive set of constraints is imposed, it can sometimes produce a simpler result than can factor analysis.

Smallest Space Analyses were carried out using as input the matrix of correlations from which the Unusual-High scores were excluded. The solution obtained in a two-vector space is illustrated in Figure 1. Vector 1 is clearly a contrast between the quality scores, all of which have negative coordinates on this axis, and the count scores, all of which have positive coordinates. This separation is equivalent to that obtained in the factor analyses; Smallest Space Analysis represents by maximum distance between two scores what factor analysis shows by having them load on different factors.

Insert Figure 1 about here

It is of interest to examine which scores provide the extreme coordinates on this vector. Among the quality scores, the Mean Quality from each test is given a higher negative coordinate than the other scores, and three of the four Mean Quality scores receive approximately the maximum extremity possible. Among the count scores, the highest positive coordinates tend to go to Unusual scores. These differences in placement are not a function of differences in the reliabilities of the several scores. The Number score for each test has higher reliability than the corresponding Unusual score, for example, yet in three of four cases it has a less extreme locus; likewise, the Highest Quality score on FH is one of the most reliable of the quality scores, yet has the second lowest coordinate of all the quality scores on this vector.

Perhaps Vector 1 of the configuration should be described as indicating that the Best and Highest Quality scores are somewhat less independent of Number than are other combinations across the quantity--quality distinction. Such a result could be artifactual: Suppose that candidates were to generate responses by selecting randomly from among the categories for a given item. Then, the more responses given, the higher would be the expected quality of the response to which we assign the highest scale value. If we credit the candidate with some ability to distinguish merit among the various responses he has generated, then the quality of the response he designates as best would also be expected to increase with the number from among which he can choose. It should be noted that, if an artifact of this sort is operating in the data, its effects are small; the correlations of Number with the Best and Highest Quality scores in Table 11 are all very low. This argument does, however, provide a conceptual basis for preferring the Mean Quality score to represent the quality domain.

For the quality scores, the second vector in the Smallest Space Analysis provides a confirmation of the factor analytic separation of EP and SMP from FH and MC. Scores from the first pair of tests receive positive coordinates, while those from the second pair receive negative ones. Moreover, the suggestion from the five-factor solution of a distinction between FH and MC is given some support. The three FH quality scores have very similar coordinates to one another, as do the three MC scores; all the former are given more extreme locations than are the latter. There is no such organization, either test- or score-specific, within the area occupied by the various EP and SMP scores.

The Vector 2 ordering found in the count scores is not the same as that among the quality scores. Here, a sharp contrast is found between Unusual scores on EP and on SMP, which are the two most closely clustered tasks so far as quality is concerned.

In summary, Smallest Space Analysis confirms the major factor analytic results--a separation of count from quality scores, and among the latter, a distinction between those from EP and SMP, on the one hand, and those from FH and MC. In addition it shows some distinction between the latter two sets of quality scores, and suggests a further and different contrast within the Unusual scores.

Relations with Background Information Questions

The Background Information Questions are shown (in somewhat abbreviated form) in Table 3. They elicit information about the candidate's educational status and training and about his plans for graduate work and career. The options are not ordered in most of the questions, and, as shown in Table 3, the response frequencies tend to be quite uneven.

A series of analyses of variance was planned to study relationships among the background information data, GRE scores, and scores on the Tests of Scientific Thinking. Before undertaking these analyses, however, a correlational study was made to determine whether the data warranted the more elaborate examination. For these analyses, a subset of the questionnaire items was selected, and dichotomies were chosen, by eliminating some response categories and combining others, to represent potentially important contrasts in the background information.

The items selected were A, B, E, F, G, and I. The nature of the dichotomies employed is indicated by the labels in Table 19. For Question A, option 2 ("I am an undergraduate senior") was contrasted with options 3, 4, and 5 combined. These three options indicated that the respondent had graduated from college and may have been in graduate school. Option 1, which was omitted, was marked by only 2.2 percent of our sample, as is shown in Table 3.

Insert Table 19 about here

For Question B, options 3 and 4, indicating that the candidate planned to obtain either a terminal master's degree or an intermediate master's on the way to the doctorate, were contrasted with option 5, indicating plans to work directly for a doctorate. Question E concerns the respondent's undergraduate major. Option 1, psychology, was contrasted with all other choices. Question F, dealing with the area of psychology in which most course work had been taken, was used to provide a sharp contrast between two possible extremes--option 1, clinical or abnormal, vs. option 3, experimental. Question G provided another sharpened contrast, that between students who had had no course work in statistics (options 1, 2, and 3) and those who had taken a statistics course (option 4). Finally, Question I dealt with the area of psychology in which the student planned to make a career. Option 2, clinical or abnormal, was contrasted with options 3 and 4, experimental or social.

The intercorrelations of the six Background Information Questions are shown in Table 19, along with their means and standard deviations. A mean of 1.5 would indicate that the candidates were evenly split between the two categories of an item, a mean of _1_ that all were in category 1 (the first category mentioned in the label) and a mean of _2_ that all were in the second category. The most uneven split was for Question E, where 88 percent of the candidates were psychology majors.

The highest correlation, .48, is between Questions F and I; it shows that students whose undergraduate work emphasized clinical and abnormal psychology also planned to make a career in clinical or abnormal psychology. The correlation of .31 between Questions A and E indicates that students who were undergraduate seniors at the time of the testing were especially likely to be psychology majors; that is, those who were taking the Advanced Psychology examination at the modal time in their training were also those whose undergraduate background was most likely to be typical. Most of the correlations in Table 19, however, are near zero, showing that for the most part we are dealing with independent items of information.

The correlations with GRE scores are shown in Table 20. The lack of correlation between Question A and ability test scores shows that students who were ahead academically had no advantage; possibly those applying after graduation or already in graduate school were individuals who were less strongly motivated, who had been admitted to a second-rate graduate school and wanted to change, or who were otherwise adjusting unsatisfactorily in their pursuit of psychological training.

Insert Table 20 about here

The highest correlations are those involving Question B. Students planning for a direct Ph.D. rather than a Master's degree were somewhat more able as measured by all the GRE scores. Students who had taken statistics (which may merely be an indication of a generally better background of undergraduate training) also tended to be the more able students, as shown for Question G. Finally, those with training in experimental rather than clinical psychology (Question F) tended to earn somewhat higher GRE scores, especially on the Experimental Psychology subtest.

Correlations with the 23 scores on the experimental tests are shown in Table 21. The table gives us little basis for improving our understanding of the new tests, since most of the correlations are very low. Only five correlations are as high as .2. Four of the five involve SMP and Questions F and I; in all four of these instances better performance on SMP is associated with emphasis on experimental rather than clinical psychology in previous training or career plans. Thus another bit of evidence is found suggesting that performance on SMP is somewhat more dependent on formal training in "hard" science than is the case for the other Tests of Scientific Thinking. The new experimental tests are again found to possess a relatively large proportion of true variance that is not attributable to identifiable aspects of formal training.

Insert Table 21 about here

On the basis of the correlations between dichotomies formed from the information items and other variables, it did not seem profitable to carry out a more detailed investigation by analysis of variance procedures.

Reanalysis for "Select" Sample

It is apparent that the group of GRE candidates as a whole includes many individuals who are only marginally qualified by training and ability or whose questionnaire responses show that they are atypical with regard to the paths they have taken in preparing for graduate training. It is of course appropriate to include all these individuals in the sample, since they were all, so far as we know, bona fide candidates for admission to graduate school, and since patterns of relationships among variables will be clearer when a wide range of talent is represented. However, it would also be of interest to examine the data for only those candidates who appear to be most appropriately trained for graduate work in psychology, who are reasonably well qualified, as judged by GRE test scores, and whose preparation for graduate work over a period of time seems to reveal consistent planning for a possible career in psychology. Such students are more likely to represent the students seriously considered as candidates by the best graduate schools and who are most likely to become graduate students.

A subsample of candidates was accordingly selected from the reliability sample on the basis of their responses to the questionnaire and their scores on the GRE Verbal Aptitude Test. Specifically, the candidates chosen were those who at the time of taking the test (1) were seniors, (2) were majoring in psychology, (3) had training in statistics as well as experimental and general psychology, and (4) earned scores of 510 or higher on the verbal aptitude test. We know from the frequency distributions that most of these students were also candidates for the Ph.D. degree. This subsample, which included 32 percent of the reliability sample, will for convenience be designated the "select" sample.

The major parts of the data analysis were repeated for the members of the select sample. Means, standard deviations, test reliabilities, and test correlations with GRE scores were all computed in order to compare the select sample with the total reliability sample.

Table 22 shows for the select sample the mean and standard deviation of each score on each of the four tests, as well as the N for each cell of the table. A comparison with Table 7 shows that in almost all instances the means are higher for the select sample than for the complete sample (which of course contains the subsample). The standard deviations are usually smaller for the select sample; the differences here are generally small except for the FH quality scores, where the ratio of the differences in standard deviations is nearly 3/2. The inconsistencies in the trend, both for means and sigmas, generally involve the scores that we know are least reliable. The students we have designated as "select" thus tended to earn slightly higher scores on the Tests of Scientific Thinking and to be less variable, which is consistent with expectations for such a group.

Insert Table 22 about here

The reliabilities for the select sample are shown in Table 23, which should be compared with Table 6. It would be expected that all correlations involving scores for the select group, including reliabilities, would be somewhat reduced because of restriction of range. The effect of restriction ought to be greatest for FH quality scores, since the differences in standard deviations are greatest for these scores. But because of the smaller N's the reliabilities for the select group can also be expected to be less stable, and some of the fluctuations are undoubtedly attributable to error.

Insert Table 23 about here

In the case of the lower-bound estimates, there seems to be no consistent difference between the select group and the complete reliability sample; differences in both directions occur with about equal frequency. There is some tendency for SMP reliabilities to be higher for the select group, even though the sigmas tend to be slightly lower. It is possible that SMP is not measuring the same abilities for the poorest candidates as for the select students because the test is too difficult for the former group. For FH quality scores reliabilities are lower for the select group, presumably because of less variability.

The upper-bound estimates of reliability for some reason tend to be higher for the select group. The general conclusion is that reliability is about as good for the select sample as for a larger sample that is representative of GRE candidates in general.

The select group was also compared with the complete reliability sample with respect to the standard errors of measurement. It was found that in 20 of the 23 comparisons the error of measurement is smaller for the select sample. The largest differences are for quality scores, especially Highest quality. For FH, for example, the standard errors of measurement for the reliability and the select samples are respectively 1.72 and 1.36. Thus the accuracy of measurement is actually higher for the select sample.

Correlations of scores with the GRE tests are shown in Table 24. Comparison of Table 24 with Table 15 shows that the correlations are generally lower for the qualified sample. Only three or four of the coefficients are appreciably higher for the more select students, but many are much lower. The median of the correlations with GRE-Verbal is .28 for the reliability sample and .13 for the select sample. For the other GRE tests the two medians are .28 and .16 for GRE-Q, .24 and .11 for Advanced Psychology, .24 and .13 for the experimental subtest, and .24 and .09 for the social subtest. Such reductions may be accounted for on the basis of the restriction in range of verbal ability.

Insert Table 24 about here

However, only in the case of GRE-V was there direct selection, producing a truncation of the distribution of GRE-V scores. For all the other tests, the selection was indirect, resulting from their correlations with GRE-V. It is not clear why this selction did not reduce correlations with the verbal test more than the others. One speculation is that the select sample contained a larger proportion of candidates who really understood the problems in the Tests of Scientific Thinking and were able to cope with them at an appropriate level, as contrasted with those who tried to deal with the problems by remembering bits of jargon or otherwise responding inappropriately.

The results for the select sample tend to verify those obtained for the reliability sample. The new tests do not duplicate the existing GRE tests with regard to abilities measured, and there is a substantial amount of true variance that is not predictable by the GRE battery. For the select sample the variance in experimental tests that is predictable from GRE tests is appreciably smaller than for the complete sample, although reliabilities in general are relatively unaffected by the restriction in range of ability.

Reanalysis for Effects of Item Order

Differences between the reliability and intercorrelation samples in the means and standard deviations of scores on the test items have been attributed to differences between these two groups in the time available for responding to each item; several methods for adjusting test variance estimates were examined in an effort to compensate for these differences. If it is assumed that candidates worked through the test items in order, it may be that the effects of speededness in the reliability sample data are largely confined to the third (and last) item in an individual's test booklet. If so, an improvement in the various test statistics might be made by analyzing data only for the first two items given to each candidate.

Test means, standard deviations, reliabilities, and correlations with GRE scores were obtained from the reliability sample data, using data for the first two of the three items administered to each candidate. In most cases (18 out of 23) higher mean test scores were found, suggesting that candidates did devote somewhat less time to the last item than to the first two. However, the differences were very slight; the largest change in means was that for the Number score, which over the four tests averaged an increase of .09 responses per item. This increase is less than one-fourth the difference between the reliability sample and the intercorrelation sample in means for the Number score; thus, the use of the first two items alone is not sufficient to eliminate context differences.

Estimated test variances showed no systematic difference between the full reliability sample data and this reduced data set. The variances were higher for 11 scores and lower for 12 in the reanalyzed data.

Correlations with the GRE scores also did not differ systematically. Over all scores, the median correlation with GRE-V was .27 in the reduced data set, as compared with .28 in the complete reliability sample. For GRE-Q the corresponding medians were .28 and .28; for the Advanced Psychology Test, .29 and .24; for the experimental subscore, .26 and .24; and for the social subscore, they were .24 and .24.

There were, on the other hand, systematic differences in the estimated reliabilities, with the reduced data set yielding somewhat higher coefficients. For the lower bound estimates, 15 of 23 reliabilities were higher; the median difference in the coefficients was .02. For the upper bound, 19 of 23 were higher, and the median difference was .09.

These results suggest a slight increase in consistency in performance when an attempt is made to control for inequalities in the candidate's allocation of time to items. However, this improvement may be partly or wholly artifactual, in that the reliability estimates may actually be inflated by the increased variability (due to smaller effective sample size) of the covariance terms entering the computation.[6] In any case, the reanalyzed data do not provide very different results from those derived from the full reliability sample; and, as they yield very similar estimates of total test means and variances, they do not afford a basis for improving estimates of test intercorrelations.

The University of Washington Study

Prior to the GRE testing, a small methodological study was conducted at the University of Washington, using as subjects 50 graduate students in psychology and educational psychology.[7] One purpose of this substudy was to obtain a set of protocols for use in developing the procedures for creating sets of categories and for using these in coding and scoring. Each of these steps could be carried out with a small block of data before attempting to employ it on protocols from 4,000 students.

A second purpose was to compare two kinds of instructions to examinees in order to answer this question: Would candidates produce solutions of higher quality if they were asked to produce the one best answer they could think of, rather than a number of solutions to each problem? Previous pilot studies had indicated that most examinees believed that the time allowed for each problem was adequate or more than adequate. However, it is possible that better performance would result if a subject could devote his whole effort to producing just one good solution to each problem; if this were true, the multiple-response instructions we had been using in pilot work might not succeed in eliciting the best problem-solving efforts of which the individual was capable. On the other hand, it could be argued that better solutions might result if the individual were asked explicitly to consider several competing solutions instead of only the one that he first believes to be the best.

A third purpose was to compare category scoring with subjective ratings of quality by scorers. Work with early versions of FH tests given to undergraduate subjects showed that the correlation between scores obtained by the two methods approached unity when corrected for unreliability. We wished to see if similar results would be found with FH test items similar to those used in the GRE study.

From a set of 18 Formulating Hypotheses problems, each subject was given six problems under instructions to write only his one best solution to each, and another six problems with instructions to write as many reasonable solutions as he could. The 18 problems were arranged in blocks of six items each. Approximately 17 students were given a block of items under each set of instructions, with order of presentation of blocks and instructions fully counterbalanced. In addition, the students were given several other cognitive ability tests and a brief questionnaire. Subjects worked for approximately one and a half hours in each of two evening sessions and were paid for their services.

The protocols from the 50 subjects were used in developing the scoring materials and procedures, including the scoring categories used in scoring the FH test. The procedures were subsequently applied to the development of materials and to the actual scoring of the protocols for all tests administered to GRE candidates.

Three scores were employed in the University of Washington study: Number of hypotheses, Mean Quality of hypotheses, and Highest quality--the quality of the best response. Data were too few to justify working with Unusual and Unusual-High scores, and in this study subjects were not asked to identify the answer they judged to be best.

Tables 25 and 26 report the intercorrelations of the three scores as obtained by both methods as well as their means and standard deviations. Table 25 shows the results under quantity instructions and Table 26 the results under quality instructions.

Insert Tables 25 and 26 about here

One of the purposes of the University of Washington substudy was to see if the high correlation between category scoring and ratings could be replicated. The correlations of interest are those underlined in Tables 25 and 26, but especially Table 25, since quantity instructions were used in the previous work and the GRE study. The correlations between category scoring and rating methods for Number were .93, .89, and .93, which are probably about as high as the score reliabilities. The three correlations for Highest quality were .46, .39, and .83. For Mean quality, the correlations were .75, .84, and .92. The analogous correlation found in the Frederiksen-Evans study was .53, although this coefficient was based on a 5-item rather than a 6-item test. Improvements in the procedures and greater care in developing categories have probably raised the correlation between Mean quality scores by the two methods, although reliabilities of scores must be taken into account in the comparison. Since *N*s for the subsamples in the present study are so small, it does not seem wise to make corrections for unreliability of the tests.

The other purpose of the substudy was to compare two kinds of instructions. Comparison of Number score means in the two tables shows that the number of hypotheses written was substantially greater under the quantity instructions, as would be expected. Under quantity instructions, subjects wrote about three hypotheses per item, and under quality instructions the means for the three blocks of items ranged from 1.3 to 1.4. (The reason the means are not *1* for quality instructions is that sometimes the single response contained more than one hypothesis, in which case both or all hypotheses were scored.) The means for the Mean quality score were consistently higher for quality instructions, which can probably be accounted for by the fact that under quantity instructions some of the added hypotheses are bound to be of lower quality.

The relevant comparison for evaluating the effects of instructions is that involving the Highest quality score. Somewhat different results are found for category scoring than for ratings. In the case of category scoring, the means are substantially higher for quantity instructions for two of the three blocks and approximately the same for the third. In the case of ratings, the differences are slight but tend to favor the quality instructions.

The results with category scoring suggest that subjects actually write better hypotheses when instructed to "write as many reasonable hypotheses as you can think of" than when asked to "write the one hypothesis which you think is most likely to explain the finding." Such an interpretation provides support for the use of quantity instructions and for obtaining both quantity and quality scores from the same protocols. From a theoretical point of view, the results suggest that when one consciously tries to find multiple solutions to a problem, even though he thinks he has the right answer, better solutions will tend to be found.

The fact that differences favor quantity instructions for categorizing but not rating requires some explanation. Perhaps the most reasonable hypothesis is that protocols written under quality instructions tend to be longer, and raters are likely to be influenced by the amount of writing. There is evidence that graders of essay questions tend to give higher grades to long answers. It is also possible that subjects who had more time for each response were able to exercise more care in writing as opposed to the more telegraphic style often used under quantity instructions. The category scoring method presumably made possible an evaluation of the basic idea expressed, apart from the length and style of its exposition.

However, another hypothesis that might account for the higher means under quantity instructions is that we have capitalized on whatever error is involved in the scoring procedure. To the extent that there is fluctuation in scores attributable to error and the highest score is chosen, we are undoubtedly capitalizing on chance variation. Such capitalization on chance should, however, affect the results for ratings as well as for category scoring. There is thus some support for the hypothesis that encouragement to produce many ideas will improve the quality of the best idea, but the problem requires further investigation.

On balance, it seems reasonable to conclude that quantity instructions are justifiable. They lead to performance in which the best response given is at least as good as, and may be better than, that obtained under quality instructions; and they permit assessment of the number and diversity of the individual's ideas in addition to indices of response quality. Moreover, results of the present investigation provide support for the use of a category-based scoring system. Category scores have very substantial correlations with those obtained by rating; and the one difference which was found between the two methods is possibly accounted for by a lesser sensitivity of these scores to confounding influences of length and style of response.

Medical School Study

A small study being conducted at the Johns Hopkins School of Medicine is of interest because it provides information based on the administration of a 6-item FH test. This information can be compared with that obtained through item-sampling methods. The study is not a part of the GRE research but was undertaken, with ETS support, partly to further the development of tests such as the Tests of Scientific Thinking.

The purpose of the study, so far as the medical school is concerned, is to try out a variety of instruments that might be useful in improving the selection of students. FH was of interest because of the obvious relevance of problem solving to medical practice. A 6-item FH test was chosen from the 18 items used in the University of Washington study. Those items were selected that appeared to be most appropriate for medical school students because of their biological flavor. Two of these items were also used in the GRE version of FH.

The FH test was administered on a voluntary basis to entering medical school students in September 1974, along with a number of other tests. Eighty students participated, but complete data were not available for all students. The battery included a long biology test, a "logical and critical reasoning" test, and two cognitive style measures, the Group Embedded Figures Test and a Draw-a-Figure test. These last two tests are used to measure field dependence-field independence; for theoretical reasons one would expect this cognitive style to be related to creative problem solving. Thus the data will extend our "nomological network" of relationships involving the Tests of Scientific Thinking.

The lower-bound estimates of reliability of the six FH scores as obtained from the medical school sample and the GRE sample are shown in Table 27. Except for Best and Unusual-High scores, where reliabilities are substantially lower for the medical school students, the reliabilities are surprisingly similar, considering the small medical school sample, the fact that the two FH tests have only two items in common, and the probability that the medical school sample is more highly selected.

Insert Table 27 about here

A comparison of the two sets of intercorrelations is shown in Table 28. Since the intercorrelations for the medical school are based on administration of a complete test, the comparable GRE figures are taken from Table 8, which shows the intercorrelations that reflect experimental dependence due to obtaining all scores from the same protocols. The correlations are roughly comparable, with a few exceptions, in spite of all the differences in test and samples.

Insert Table 28 about here

Table 28 also shows the means for the two groups. The quality scores are not comparable because the units of measurement have different meaning for different items. The medical school students wrote more hypotheses, gave about the same number of unusual responses, and had fewer which were unusual-high quality.

By and large, the comparison of reliabilities, intercorrelations, and means suggests that the data obtained through item-sampling methods result in scores whose characteristics are quite similar to those produced by the administration of an entire test.

Summary

The purpose of the investigation was to develop a set of tests that might reasonably be used as provisional criterion measures in research on scientific thinking, particularly "creative" thinking; and to assess the suitability of these tests as criterion variables from the standpoint of their psychometric properties. The properties to be investigated include test difficulty, reliability, intercorrelations, and correlations with other variables. The tests would be useful in research on creativity and perhaps also for other purposes such as individual assessment, selection, and training evaluation, if it could be shown that they are of suitable difficulty for mature students, that they are reliable, and that they are valid. In the absence of acceptable external criteria, the validation must consist of demonstrations that the tests do not merely reflect the abilities already measured by conventional test batteries (discriminant validity) and that they do correlate with other measures in a way that is consistent with theoretical expectations (convergent validity).

The four tests that were developed are performance tests that simulate aspects of the job of a behavioral scientist; they are Formulating Hypotheses (FH), Evaluating Proposals (EP), Solving Methodological Problems (SMP), and Measuring Constructs (MC). Each item poses a problem such as a research psychologist might face, and the task of the examinee is to propose a number of solutions--not only the one he considers best, but also others that he thinks ought to be considered.

Since it might be argued that asking a candidate to write several answers would result in solutions of lower quality than asking him to

write the one best answer he can think of, a small study was carried out with 50 graduate students at the University of Washington to compare the two kinds of instructions. Each subject was given six FH items under each kind of instruction. The crucial comparison involves that of the average quality of responses under the one-answer instruction with the best responses under the multiple-answer condition. The results suggest that the quality of the best response is actually higher under the multiple-response instruction (although there is some possibility that the difference results from capitalization on error). The finding indicates that the kind of instructions we have used does not handicap the examinee in writing answers of high quality.

A scoring system was developed in which, for each item, the scorer is given a list of categories which includes almost all of the ideas that examinees write in response to that problem. Rather than making subjective judgments of response quality, the scorer has only to assign each response to the appropriate category. The categories in each list have independently been evaluated and assigned scale values from which quality scores can be derived. This method is judged to be faster and more accurate than direct ratings of quality, and may produce scores which are less influenced by such extraneous factors as neatness and legibility of handwriting. A comparison of scores obtained by rating and by category scoring methods shows high agreement relative to the reliabilities of the scores. Correlations range up to .93 for the most reliable of the quality scores.

In addition to obtaining a quality scale value for each response, the number of responses given by each examinee to each problem is determined.

Six scores for each test can then be generated by computer: (1) Best, the average quality of the responses designated by the examinee as the best for each item; (2) Mean Quality, the average quality over all the responses; (3) Highest, the mean quality of the responses that were best according to the category scoring method; (4) Number, the number of non-duplicate responses; (5) Unusual, the number of responses in categories that occurred rarely; and (6) Unusual-High, the number of responses that were both unusual and of high quality. The last score is analogous to a common definition of a creative product: one that is both novel and useful.

The four 6-item tests were administered to about 4,000 candidates taking the GRE Advanced Psychology Test, using an item-sampling method. Each candidate was given a subtest containing either three items from one test or two items from different tests. These subtests represented all the possible combinations of items in all possible orders, and they were administered to randomly-designated subgroups of students. This procedure made it possible to compute all the item variances and co-variances, from which it was possible to compute for each score the test reliability, correlations with other scores on the experimental tests, and relations of tests to other variables.

Lower-bound estimates of the reliabilities of the scores for 6-item tests vary widely, with the Number score being the most reliable (ranging from .61 for SMP to .77 for MC). Mean Quality is the most reliable of the quality scores, having a range from .46 for SMP to .77 for MC, with Highest quality a close second. Unusual and Unusual-High are less reliable than Number, being based on much smaller subsets of responses, and Unusual-High is too unreliable to be useful for SMP and MC. SMP is consistently the least reliable of the four tests. With the exceptions noted, reliabilities are high enough to make the tests useful. The reliability of any test can of course be increased by adding more items.

The number of responses to an item is typically about three or four, while the number of Unusual or Unusual-High responses is much lower, which accounts for their lower reliability. The Mean Quality and the other quality scores are well up in the range of possible scale values. Thus the items seem to be of appropriate difficulty for graduate school applicants and probably also for more advanced students or even junior faculty members. The SMP may have been a little too difficult for the GRE candidates but may be appropriate for more advanced students. Reliabilities were found to be approximately the same for a more select subgroup of the examinees, in spite of the smaller range of talent, which supports the conclusion that the tests are appropriate for examinees at a high level of training.

The intercorrelations show that within a test the three quality scores form a cluster of related variables, as do the three scores based on number of responses (although low reliabilities of the Unusual-High scores make the interpretation less clear). A factor analysis of the score intercorrelations reveals two quality factors and one number-of-responses factor. The two quality factors are defined by tests: one factor reflects quality scores for FH and MC; the other, quality scores for EP and SMP. The quantity factor involves quantity scores (Number and Unusual) for all four tests. A nonmetric Smallest Space Analysis produces a similar result. Thus it is clear that performance on the tests can grossly be described in terms of two major dimensions--number of ideas and quality of ideas--and that there are at least two kinds of quality involved. An interesting problem for the future will be to find out just what the distinction is between the two kinds of quality and to see if other dimensions of quality can be identified. It is also clear that it would be unwise to report a single measure of quality based on all four tests, although a Number score based on all four might be justified.

Correlations with the GRE tests are relatively small, even when corrected for unreliability. Quality scores are more closely related to GRE scores than are the count scores, and the correlations are highest for SMP. But there is a substantial amount of true variance in the Tests of Scientific Thinking that is not accounted for by GRE scores; typically only about 25% of the true variance is accounted for by any GRE test. Thus there is evidence that the new tests have sufficient discriminant validity to be potentially useful.

The correlations with GRE scores contribute little to construct validation, partly because the correlations are not high and partly because the GRE scores are related nearly equally to all the scores and all the tests. Presumably the reason is that GRE test intercorrelations are high--they tend to measure similar abilities.

Six dichotomous scores based on a set of Background Information items were also correlated with the experimental test scores. These items dealt with amount and kind of training and plans for further education and career. These correlations were almost uniformly low. In most cases the background information items also failed to correlate substantially with GRE scores; thus, they make little contribution to construct validation.

What can be concluded about the Tests of Scientific Thinking? Might they be useful as criterion measures in investigations of processes involved in creative problem solving? The needed evidence has to do with test difficulty, test reliability, and test validity.

With regard to difficulty, the tests do seem to be sufficiently challenging for graduate students, and perhaps even for research scientists. We found that even for a select subsample of applicants for admission to graduate school the tests are by no means too easy, and the accuracy of measurement is actually greater when the marginal candidates are eliminated.

Scores derived from each of the tests form two clusters, one reflecting quality, the other quantity, of ideas. For the quality scores and for a count of number of responses, reliabilities for a 6-item test are quite adequate for a research instrument. The two remaining scores--number of Unusual responses, which falls into the quantity cluster, and number of Unusual-High Quality responses, which does not have consistent relations to other scores--have lower reliabilities, especially for two of the tests.

Since a 6-item test can be given in less than 50 minutes, it would be feasible to employ still longer tests, and consequently to increase the reliabilities of all the scores.

Evidence on validity shows that the tests do not merely reflect abilities already measured by the GRE aptitude and achievement tests. The scores based on number of ideas and number of unusual ideas in particular are unrelated to the conventional tests. The quality scores have some correlation with GRE measures, as would be expected on theoretical grounds, but even if all the tests could be made perfectly reliable, only a small proportion of the variability in quality scores would be predictable from GRE scores. Thus the Tests of Scientific Thinking do have reliable variance that is not being measured by existing tests.

Another consideration has to do with "face" validity: do the tests appear to be capable of measuring important aspects of problem solving to one who has examined them carefully and actually taken the tests? Our opinion is that they do, based on comments of scientists and educators who have examined the tests and also on occasional written comments by students who were our subjects.

Further investigation seems warranted. A follow-up study of our GRE sample is now in progress. Other studies should be directed primarily at questions of construct validity: is performance on the tests related to other personal characteristics in ways that are theoretically consistent, and does performance change in relation to experimental or educational treatments in directions that would be expected on logical grounds? Answers to such questions will not only provide evidence on validity but will also contribute to an understanding of processes involved in scientific thinking.

References

Barron, F. The psychology of creativity. Chapter in New directions in psychology II. New York: Holt, Rinehart and Winston, 1965.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Flanagan, J. C., et al. Critical requirements for research personnel: A study of observed behavior of personnel in research laboratories. Pittsburgh: American Institute for Research, 1949.

Frederiksen, N., & Evans, F. R. Effects of models of creative performance on ability to formulate hypotheses. Journal of Educational Psychology, 1974, 66, 67-82.

Guilford, J. P. Some new looks at the nature of creative processes. In N. Frederiksen and H. Gulliksen (Eds.), Contributions to mathematical psychology. New York: Holt, Rinehart and Winston, 1964.

Guilford, J. P. The nature of human intelligence. New York: McGraw-Hill, 1967.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Guttman, L. A basis for analyzing test-retest reliability. Psychometrika, 1945, 10, 255-282.

Guttman, L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika, 1968, 33, 469-506.

Jöreskog, K. G., Gruvaeus, G. T., & van Thillo, M. ACOVS: A general computer program for analysis of covariance structures. Research Bulletin 70-15. Princeton, N. J.: Educational Testing Service, 1970.

Koutsopoulos, C. J. The mathematical foundations of classical test theory: An axiomatic approach II. Research Memorandum 64-3. Princeton, N. J.: Educational Testing Service, 1964.

MacKinnon, D. W. The nature and nurture of creative talent. American Psychologist, 1962, 17, 484-495.

Ward, W. C. Rate and uniqueness in children's creative responding. Child Development, 1969, 40, 869-878.

Footnotes

[1] This research was supported by the Graduate Record Examinations Board.

[2] The four judges were the two authors and two Research Assistants, Suzanne Taweel and Charlotte Kiefer, who not only aided in the development of the scoring procedure but also supervised the scoring and contributed to the research in many other ways.

[3] Robert Altman and Martin Glaubitz were helpful in making arrangements for the inclusion of the experimental tests.

[4] Frederic Lord provided useful advice in developing the procedures necessary for these computations.

[5] Ledyard Tucker suggested several alternative procedures for the factor analysis.

[6] The reliability estimates may be more susceptible than the other statistics to the reduction in amount of data. The number of candidates providing data for any one item is reduced by one-third by deleting the last item given, but _N_'s for the covariance terms used in the reliability computation are reduced by two-thirds, since each candidate contributes to one rather than three such terms. Moreover, the lower bound estimates involve a term obtained by summing squared item covariances; with a decrease in sample size for each term, the sampling distribution will increase in variability, leading to occasional overly large terms which will make unreasonably large contributions to the estimate. Similarly,

the upper bound estimates involve selecting the largest of 10 split-half coefficients; as the distribution of these coefficients becomes more variable, the probability of selecting one which is unreasonably large increases.

[7] Patricia Cox was responsible for collecting data at the University of Washington.

Table 1

Alpha Coefficients Showing Amount of Agreement Among Four Judges in Ranking Categories in Order of Quality

Test	Item					
	1	2	3	4	5	6
Formulating Hypotheses	.98	.89	.98	.94	.96	.97
Evaluating Proposals	.92	.96	.92	.83	.91	.87
Solving Methodological Problems	.76[ab]	.91	.91	.87	.84	.92
Measuring Constructs	.79[b]	.89[ab]	.94	.77[ab]	.95	.95

[a] Only three judges ranked the categories for these items.

[b] For these items there was a much larger number of categories because most responses could be classified in two ways (see text). For these items, only those categories used by candidates were judged, and they were rated on a 21-point scale rather than ranked.

Table 2

Sample for GRE Administration Study

Total domestic candidates		4,394
Test booklets not returned (presumably because candidates were given an alternate form of the Advanced Psychology Test that did not contain the experimental test)	64	
Booklets not scorable (blank, overt refusal, or no permission signature)	373	
Discarded (data from two centers that were not cooperative)	26	
Administrative errors (e.g., could not be matched with GRE scores)	20	
Incomplete data (e.g., last item not completed)	325	
Total candidates dropped		808
Complete data (used in analysis)		3,586

Table 3

Responses to Background Information Questionnaire

Response	% of October 1973 Group	% of Sample
A. At what point are you in your studies?		
0. No response	1.8	1.1
1. In or just completed junior year	2.1	2.2
2. Senior	65.2	68.3
3. AB, not in graduate school	16.1	15.4
4. In, or just completed, first year of graduate school	8.0	7.7
5. In, or just completed, second year of graduate school	6.7	5.4
B. What graduate degree do you intend to seek?		
0. No response	2.0	1.2
1. Do not plan graduate study	1.4	1.2
2. Plan graduate work but no degree	.6	.4
3. Terminal master's degree	16.1	15.4
4. Master's degree leading to doctorate	24.6	25.1
5. Doctoral degree	55.3	56.6
C. If you are a senior, which best describes your education and educational plans?		
0. No response	7.3	6.1
1. Psychology major, planning graduate work in psychology	54.5	56.8
2. Psychology major, planning graduate work in related field	8.0	8.0
3. Psychology major, planning graduate work in different field	1.2	1.2
4. Not psychology major, planning graduate work in psychology	2.4	2.3
5. Not a senior; other	26.7	25.5
D. In what general area would you classify your undergraduate major?		
0. No response	2.6	1.8
1. Social science	81.6	83.0
2. Biological science	4.2	4.5
3. Physical science	.9	.9
4. Mathematics	.5	.5
5. Other	10.2	9.3

Table 3 (Continued)

Response	% of October 1973 Group	% of Sample
E. What is(was) your undergraduate major?		
0. No response	2.5	1.9
1. Psychology	85.8	87.7
2. Philosophy	1.0	.9
3. Sociology	1.3	1.1
4. Education	1.5	1.2
5. Other	7.9	7.2
F. In what area of psychology have you had the most course work?		
0. No response	2.9	2.0
1. Clinical or abnormal	34.0	33.6
2. Educational	5.8	5.5
3. Experimental	27.6	29.4
4. Social	15.6	15.9
5. Other	14.2	13.6
G. Which of the following best describes your work?		
0. No response	2.3	1.5
1. General psychology only	10.5	9.7
2. Experimental psychology only	1.2	1.2
3. General and experimental psychology only	9.6	9.3
4. General psychology, experimental psychology, and statistics	67.4	69.6
5. Other	9.0	8.7
H. How recently have you had a college or graduate course in psychology?		
0. No response	2.1	1.1
1. During current academic year	75.2	77.3
2. During previous academic year	15.8	15.3
3. Two or three years ago	4.7	4.5
4. Four or five years ago	1.1	.9
5. Other	1.1	.9
I. In what area of psychology do you plan your career?		
0. No response	2.8	1.8
1. Clinical or abnormal	51.1	51.3
2. Educational	9.7	9.8
3. Experimental	11.1	11.4
4. Social	8.3	8.5
5. Other, or not in psychology	17.0	17.1

Table 4

Comparison of Sample with all GRE Candidates

on Test Performance

	N		Mean		SD	
	Oct. 73 Group	Sample	Oct. 73 Group	Sample	Oct. 73 Group	Sample
GRE Verbal	3709	2998	551	558	103	100
GRE Quantitative	3709	2998	540	549	120	118
Advanced Psychology	4469	3560	546	552	94	93
Subtest 1: Experimental	4469	3560	54	55	10	9
Subtest 2: Social	4469	3560	54	55	9	9

Table 5

Interscorer Reliability by Item
(Coefficient Alpha)

Test	Item	Score					
		Best	Mean Quality	Highest Quality	Number	Unusual	Unusual-High
FH	1	.51	.68	.61	.90	.72	.71
	2	.79	.80	.82	.93	.84	.82
	3	.60	.74	.77	.88	.80	.87
	4	.86	.87	.89	.90	.76	.79
	5	.86	.88	.87	.90	.81	.78
	6	.73	.76	.69	.88	.64	.66
	Median	.76	.78	.79	.90	.78	.79
EP	1	[a]	.78	.76	.94	.77	.72
	2		.82	.82	.91	.72	.61
	3		.80	.87	.92	.79	.75
	4		.86	.89	.96	.91	.81
	5		.80	.83	.94	.77	.68
	6		.73	.81	.92	.79	.82
	Median		.80	.83	.93	.78	.73
SMP	1	.76	.76	.78	.90	.57	.38
	2	.85	.89	.90	.93	.69	.72
	3	.82	.85	.86	.89	.69	.21
	4	.62	.80	.85	.86	.63	.58
	5	.77	.81	.77	.90	.71	.64
	6	.88	.90	.93	.88	.66	.61
	Median	.79	.83	.85	.89	.67	.59
MC	1	.68	.62	.69	.86	.68	.57
	2	.81	.90	.86	.90	.70	.71
	3	.79	.76	.77	.89	.78	.73
	4	.77	.83	.83	.87	.61	.46
	5	.77	.82	.86	.86	.76	.66
	6	.67	.77	.77	.87	.63	.60
	Median	.77	.79	.80	.87	.69	.63
Over all tests		.77	.80	.83	.90	.72	.69

[a] Candidates were not asked to choose their best answer for EP.

Table 6

Reliabilities of Scores on the Tests of Scientific Thinking[a]

Test		Score					
		Best	Mean Quality	Highest	Number	Unusual	Unusual-High
FH	lower bound	.50	.62	.65	.67	.53	.46
	upper bound	.62	.71	.79	.78	.69	.59
EP	lower bound		.61	.51	.73	.55	.44
	upper bound		.79	.60	.81	.74	.58
SMP	lower bound	.30	.46	.35	.61	.42	.05
	upper bound	.51	.62	.48	.73	.60	.21
MC	lower bound	.68	.77	.71	.77	.36	.17
	upper bound	.80	.88	.85	.82	.46	.34

[a] *N*'s for the reliability sample were 359, 339, 309, and 340, for FH, EP, SMP, and MC, respectively. Each of these examinees provided scores on three items from one test, except that a number of examinees sometimes omitted marking a "Best" response. Effective *N*'s for Best response are 323, 248, and 290 for FH, SMP, and MC, respectively; examinees were not asked to designate a best response on EP.

Table 7

Means and Standard Deviations of Scores on the Tests of Scientific Thinking (Based on Reliability Study Sample)[a]

Test		Score					
		Best	Mean Quality	Highest	Number	Unusual	Unusual-High
FH	Mean	19.67	17.88	22.09	2.45	.61	.24
	SD	3.42	2.97	2.91	.57	.36	.22
EP	Mean		17.80	23.94	3.90	1.10	.18
	SD		1.94	1.84	.93	.49	.18
SMP	Mean	14.89	14.00	17.18	2.42	.52	.12
	SD	2.23	1.95	1.84	.57	.29	.11
MC	Mean	14.55	13.70	17.95	2.78	.77	.21
	SD	3.64	2.96	3.20	.79	.36	.17

---

[a] See footnote to Table 6.

Table 8

Number of Categories and Theoretical Upper Limit of Scores

Test	Item	Number of Categories	Highest Possible Score		
			Quality Scores	Unusual	Unusual-High
FH	1	25	24.25	14	4
	2	34	33.25	23	8
	3	25	24.75	15	6
	4	25	23.75	13	3
	5	23	22.75	14	5
	6	24	23.75	14	4
	Mean		25.42	15.50	5.00
EP	1	25	23.50	15	2
	2	25	24.50	13	3
	3	31	30.25	20	4
	4	35	32.50	23	3
	5	26	25.25	15	2
	6	28	26.00	17	2
	Mean		27.00	17.17	2.67
SMP	1[a]	99	21.00	57[b]	14[b]
	2	23	20.75	14	4
	3	23	22.25	14	1
	4	23	21.50	13	2
	5	29	29.00	16	2
	6	21	19.25	10	3
	Mean		22.29	13.40	2.40
MC	1[a]	143	21.00	90[b]	23[b]
	2[a]	206	21.00	122[b]	37[b]
	3	29	27.50	18	4
	4[a]	148	21.00	85[b]	26[b]
	5	32	30.25	18	3
	6	31	29.50	19	5
	Mean		25.04	18.33	4.00

Footnotes for Table 8

[a] A system involving dual classification of responses was used, and categories were rated, using a 21-point scale, instead of ranked.

[b] These values were excluded in computing means.

Table 9

Means and Standard Deviations of Scores on the Tests of Scientific Thinking (Based on Intercorrelation Study Sample)[a, b]

Test		Score					
		Best	Mean Quality	Highest	Number	Unusual	Unusual-High
FH	Mean	19.40	17.44	22.23	2.84	.78	.29
	SD	3.42	2.97	2.91	.65	.44	.25
EP	Mean		17.75	24.33	4.36	1.22	.21
	SD		1.94	1.84	1.03	.53	.20
SMP	Mean	14.75	13.44	17.25	2.78	.60	.14
	SD	2.23	1.95	1.84	.64	.33	.13
MC	Mean	14.10	13.34	18.28	3.26	.85	.22
	SD	3.64	2.96	3.20	.91	.39	.18

[a] Ns for the intercorrelation sample were 1134, 1120, 1115, and 1109, for FH, EP, SMP, and MC, respectively. A number of examinees sometimes omitted marking a "Best" response, however, leading to effective Ns for this score of 1048, 939, and 952, for FH, SMP, and MC, respectively; examinees were not asked to designate a best response on EP.

[b] SDs of scores were taken from performance of the Reliability Study sample, after correcting variances of the count scores for the difference in test "length." See text and Appendix C.

Table 10

Intercorrelations of Scores for Each Test Using All Covariances

Test		Score				
		Mean	Highest	Number	Unusual	Unusual-High
FH	Best	.81	.76	.14	.10	.28
	Mean		.82	-.06	-.11	.20
	Highest			.35	.09	.25
	Number				.65	.46
	Unusual					.67
EP	Best	--	--	--	--	--
	Mean		.74	-.07	-.33	.10
	Highest			.47	.11	.26
	Number				.74	.36
	Unusual					.48
SMP	Best	.85	.64	-.30	-.34	.16
	Mean		.73	-.40	-.47	.13
	Highest			.16	-.20	.20
	Number				.50	.14
	Unusual					.27
MC	Best	.83	.78	.02	.10	.36
	Mean		.87	-.04	-.08	.39
	Highest			.33	.17	.46
	Number				.68	.29
	Unusual					.45

Table 11

Intercorrelations of Scores for Each Test Eliminating Experimental Dependencies

Test		Score				
		Mean	Highest	Number	Unusual	Unusual-High
FH	Best	.50	.48	.20	.20	.28
	Mean		.52	.05	.03	.15
	Highest			.31	.13	.20
	Number				.40	.31
	Unusual					.35
EP	Best	--	--	--	--	--
	Mean		.44	-.03	-.20	.01
	Highest			.35	.10	.18
	Number				.56	.26
	Unusual					.26
SMP	Best	.43	.21	-.28	-.27	-.02
	Mean		.20	-.37	-.42	-.08
	Highest			-.00	-.25	-.06
	Number				.30	.01
	Unusual					-.14
MC	Best	.64	.60	.03	.09	.26
	Mean		.65	-.03	-.11	.26
	Highest			.24	.08	.29
	Number				.49	.20
	Unusual					.03

Table 12

True Score Intercorrelations of Scores for Each Test[a]

Test		Score				
		Mean	Highest	Number	Unusual	Unusual-High
FH	Best	.75	.69	.29	.30	.46
	Mean		.70	.07	.05	.23
	Highest			.39	.17	.29
	Number				.54	.46
	Unusual					.55
EP	Best	--	--	--	--	--
	Mean		.63	-.03	-.26	.02
	Highest			.50	.15	.30
	Number				.72	.41
	Unusual					.40
SMP	Best	.76	.42	-.46	-.50	-.05
	Mean		.37	-.55	-.68	-.23
	Highest			-.01	-.47	-.18
	Number				.45	.03
	Unusual					-.41
MC	Best	.76	.73	.04	.15	.49
	Mean		.75	-.04	-.17	.47
	Highest			.29	.13	.53
	Number				.79	.38
	Unusual					.09

[a] Correction for unreliability of correlations shown in Table 11, based on "parallel form" reliability estimates.

Table 13

Correlations of Corresponding Scores

from Different Tests

	Best	Mean Quality	Highest	Number	Unusual	Unusual-High
FH-EP	--	.53	.18	.68	.43	.22
FH-SMP	.37	.06	.11	1.03	.16	.35
FH-MC	.38	.24	.35	.48	.19	-.26
EP-SMP	--	.63	.76	.42	.04	-.23
EP-MC	--	.53	.60	.30	.04	.21
SMP-MC	.34	.17	.32	.59	.42	.41

Table 14

True Score Correlations of Corresponding Scores from Different Tests[a]

	Best	Mean Quality	Highest	Number	Unusual	Unusual-High
FH-EP	--	.71	.27	.86	.60	.38
FH-SMP	.67	.08	.18	1.36	.26	.98
FH-MC	.54	.31	.42	.60	.33	-.57
EP-SMP	--	.90	1.42	.54	.06	-.66
EP-MC	--	.64	.84	.37	.06	.43
SMP-MC	.54	.23	.50	.76	.81	1.51

---

[a] Correction for unreliability of correlations shown in Table 13, based on "parallel form" reliability estimates.

Table 15

Correlations with GRE Scores

Score	Test	Verbal	Quantitative	Advanced Psychology	Experimental Subscore	Social Subscore
Best	FH	.34	.34	.28	.24	.25
	EP	--	--	--	--	--
	SMP	.39	.32	.40	.33	.42
	MC	.28	.30	.24	.24	.18
Mean	FH	.31	.28	.24	.20	.24
	EP	.37	.31	.33	.29	.31
	SMP	.39	.34	.45	.36	.46
	MC	.29	.34	.25	.24	.18
Highest	FH	.38	.37	.31	.26	.31
	EP	.47	.40	.42	.37	.39
	SMP	.49	.39	.53	.50	.46
	MC	.35	.33	.31	.25	.29
Number	FH	.25	.35	.22	.21	.17
	EP	.35	.27	.29	.27	.22
	SMP	.15	.14	.13	.19	.02
	MC	.22	.05	.19	.10	.27
Unusual	FH	.13	.20	.12	.13	.09
	EP	.13	.09	.07	.09	.01
	SMP	.01	.03	-.09	-.03	-.14
	MC	.28	.10	.23	.18	.26
Unusual-High	FH	.19	.19	.17	.14	.16
	EP	.15	.13	.11	.15	.03
	SMP	.20	.21	.18	.21	.10
	MC	.24	.21	.32	.26	.32

Table 16

True-Score Correlations with GRE Scores[a]

Score	Test	Verbal	Quantitative	Advanced Psychology	Experimental Subscore	Social Subscore
Best	FH	.44	.45	.37	.33	.35
	EP	--	--	--	--	--
	SMP	.57	.48	.59	.50	.65
	MC	.33	.36	.28	.29	.23
Mean	FH	.38	.35	.30	.26	.32
	EP	.43	.36	.38	.35	.39
	SMP	.51	.45	.60	.49	.65
	MC	.32	.35	.27	.27	.22
Highest	FH	.43	.43	.37	.32	.39
	EP	.78	.54	.57	.52	.56
	SMP	.72	.59	.81	.78	.73
	MC	.39	.38	.35	.29	.35
Number	FH	.29	.41	.25	.25	.22
	EP	.40	.32	.34	.33	.27
	SMP	.18	.17	.16	.24	.03
	MC	.25	.06	.22	.11	.33
Unusual	FH	.16	.25	.15	.16	.12
	EP	.15	.11	.09	.11	.02
	SMP	.01	.04	-.12	-.04	-.20
	MC	.42	.16	.36	.28	.43
Unusual-High	FH	.25	.26	.23	.20	.22
	EP	.20	.18	.16	.22	.04
	SMP	.44	.47	.41	.48	.23
	MC	.43	.37	.57	.49	.62

---

[a] Correction for unreliability of correlations shown in Table 15, based on "parallel form" reliability estimates.

Table 17

Varimax Factor Loadings
Using All Scores for Each Test[a]

Test	Score	Factor I	Factor II	Factor III
FH	Best	.05	.43	.27
	Mean Quality	-.15	.45	.16
	Highest	.14	.52	.06
	Number	.82	.32	.09
	Unusual	.70	.13	.05
	Unusual-High	.39	.29	.41
EP	Mean Quality	-.21	.51	.56
	Highest	.10	.37	.69
	Number	.68	.03	.35
	Unusual	.51	-.07	-.19
	Unusual-High	.06	.37	-.27
SMP	Best	.23	.27	.67
	Mean Quality	-.27	.06	.72
	Highest	.28	.08	.71
	Number	.86	-.09	-.12
	Unusual	.45	.11	-.42
	Unusual-High	.28	.29	.08
MC	Best	.02	.74	.08
	Mean Quality	-.09	.75	.10
	Highest	.14	.72	.28
	Number	.58	.03	.27
	Unusual	.53	-.22	-.03
	Unusual-High	.28	.23	.03

[a] Loadings of .40 and greater have been underlined.

Table 18

Varimax Factor Loadings and Extension Variable Loadings Deleting the Unusual-High Score for Each Test[a]

Test	Score	Factor		
		I	II	III
FH	Best	.04	.58	.09
	Mean Quality	-.14	.57	.03
	Highest	.14	.64	-.09
	Number	.80	.40	-.04
	Unusual	.68	.17	-.05
EP	Mean Quality	-.19	.53	.55
	Highest	.15	.35	.73
	Number	.70	.04	.32
	Unusual	.53	-.18	-.09
SMP	Best	.25	.45	.45
	Mean Quality	-.22	.12	.72
	Highest	.33	.14	.71
	Number	.84	-.10	-.18
	Unusual	.40	.21	-.56
MC	Best	-.01	.64	.13
	Mean Quality	-.10	.64	.18
	Highest	.15	.65	.35
	Number	.61	.06	.24
	Unusual	.55	-.16	-.04
Extension Variables	GRE-V	.27	.39	.41
	GRE-Q	.22	.44	.28
	Adv. Psychology	.23	.30	.47
	SS1	.22	.27	.38
	SS2	.18	.26	.47

[a] Loadings of .40 and greater have been underlined.

Table 19

Intercorrelations of Background Information Items

(Based on Reliability Sample)

	A	B	E	F	G	I
A. Senior vs. AB or graduate school		-.02	.31	-.14	-.11	-.04
B. Plan MA vs. Ph.D.	-.02		.02	.04	.09	-.02
E. Psychology major vs. other	.31	.02		-.12	-.19	-.06
F. Course work in clinical vs. experimental	-.14	.04	-.12		.12	.48
G. Had no statistics vs. some	-.11	.09	-.19	.12		.05
I. Plans career in clinical vs. experimental or social	-.04	-.02	-.06	.48	.05	
N	1275	1278	1291	828	1181	969
Mean	1.29	1.59	1.10	1.46	1.76	1.28
S.D.	.45	.49	.31	.50	.43	.45

Table 20

Correlations of Background Information Items with GRE Scores

(Based on Reliability Sample)

	GRE-V	GRE-Q	Adv. Psych.	Exp. Psych.	Soc. Psych.
A. Senior vs. AB or graduate school	-.05	-.08	.02	.00	.04
B. Plan MA vs. Ph.D.	.24	.25	.33	.31	.29
E. Psychology major vs. other	.03	-.01	-.03	-.08	.02
F. Course work in clinical vs. experimental	.06	.17	.14	.25	-.04
G. Had no statistics vs. some	.09	.14	.19	.16	.15
I. Plans career in clinical vs. experimental or social	.04	.11	.05	.12	-.06

Table 21

Correlations of Background Information Items with the Tests of Scientific Thinking (Based on Reliability Sample)

Score	Test	A	B	E	F	G	I
Best	FH	.04	.06	.02	.06	.04	.01
	EP	--	--	--	--	--	--
	SMP	-.09	.13	-.06	.23	.11	.13
	MC	-.05	.13	-.13	.06	.12	.08
Mean	FH	.07	-.00	.02	.03	-.02	.07
	EP	.07	.15	-.03	.13	.17	.11
	SMP	.06	.07	-.03	.17	.02	.04
	MC	-.03	.09	-.12	.08	.12	.05
Highest	FH	.02	.08	.03	-.03	.05	.02
	EP	.00	.20	-.09	.05	.15	.04
	SMP	.01	.15	-.04	.20	.10	.06
	MC	-.06	.11	-.07	-.03	.12	-.05
Number	FH	-.04	.11	.13	.15	.08	.10
	EP	-.07	.10	-.05	-.08	.05	-.10
	SMP	.06	.16	.05	-.01	.08	.02
	MC	-.00	.04	.02	-.08	-.01	-.14
Unusual	FH	.00	.13	.01	.15	.03	.02
	EP	-.05	-.06	-.08	-.08	.04	-.11
	SMP	.02	.08	-.01	-.00	-.03	.02
	MC	.04	.08	.03	-.04	.02	-.03
Unusual-High	FH	.01	.11	.11	.12	-.03	.02
	EP	-.04	-.04	-.03	.06	.01	.04
	SMP	-.06	.01	-.05	.20	-.14	.20
	MC	.02	.11	.02	.06	.11	.10

Table 22

Means and Standard Deviations of Scores on the Tests of Scientific Thinking (Based on the Select Sample)

Test		Score					
		Best	Mean Quality	Highest	Number	Unusual	Unusual-High
	N	88	96	96	96	96	96
FH	Mean	20.57	18.30	22.77	2.56	.62	.25
	S.D.	2.22	1.99	2.03	.59	.33	.20
	N		110	110	110	110	110
EP	Mean		18.07	24.22	3.98	1.13	.19
	S.D.		1.69	1.84	.85	.48	.18
	N	91	112	112	112	112	112
SMP	Mean	15.47	14.19	17.56	2.50	.50	.12
	S.D.	2.21	1.93	1.82	.59	.28	.10
	N	91	106	106	106	106	106
MC	Mean	15.44	14.54	18.78	2.77	.80	.24
	S.D.	3.36	2.84	2.66	.78	.37	.18

Table 23

Estimates of Reliabilities of the Tests of Scientific Thinking[a] (Based on the Select Sample)

Test		Score					
		Best	Mean Quality	Highest	Number	Unusual	Unusual-High
FH	lower bound	.08	.37	.55	.74	.51	.50
	upper bound	.56	.75	.74	.94	.73	.66
EP	lower bound		.50	.63	.70	.50	.50
	upper bound		.78	.83	.95	.74	.65
SMP	lower bound	.30	.56	.48	.66	.53	.02
	upper bound	.47	.90	.71	.94	.75	.02
MC	lower bound	.67	.78	.71	.77	.39	.21
	upper bound	.89	.94	.85	.89	.47	.54

---

[a] *N*'s for the select sample were 96, 110, 112, and 106 for FH, EP, SMP, and MC, respectively, except that *N*'s were somewhat lower for "Best" response, as shown in Table 22.

Table 24

Correlations with GRE Scores
(Based on the Select Sample)

Score	Test	Verbal	Quantitative	Advanced Psychology	Experimental Subtest	Social Subtest
Best	FH	-.04	.08	-.18	-.21	-.12
	EP	--	--	--	--	--
	SMP	.29	.14	.37	.30	.39
	MC	.04	.23	.07	.01	.09
Mean	FH	.05	.13	-.18	-.14	-.12
	EP	.09	-.05	.11	.09	.08
	SMP	.31	.20	.45	.37	.42
	MC	.05	.25	.08	.03	.09
Highest	FH	.14	.16	-.04	-.07	.06
	EP	.18	.08	.18	.13	.16
	SMP	.24	.17	.49	.52	.31
	MC	.06	.27	[illegible]	.03	.16
Number	FH	.24	.21	.16	.16	.15
	EP	.49	.25	.39	.30	.29
	SMP	-.03	.17	.06	.19	-.20
	MC	.12	.04	.01	-.04	.11
Unusual	FH	.16	.11	.21	.17	.19
	EP	.32	.15	.15	.16	.07
	SMP	.11	.20	-.07	-.05	-.11
	MC	.06	.12	.06	.04	.07
Unusual-High	FH	.18	.12	.18	.07	.21
	EP	.13	.01	.10	.17	-.05
	SMP	.42	.22	.40	.32	.35
	MC	-.11	.18	.05	-.02	.09

Table 25

Correlations Between FH Scores Obtained by Categorizing and by Rating Methods (Quantity Instructions)

		Categorizing			Rating			Mean	SD
		Number	Mean Quality	Highest	Number	Mean Quality	Highest		
Items 1-6 (N = 18)									
Categorizing	Number		-.01	.24	.93	-.03	.18	2.96	.55
	Mean Quality			.72	-.14	.75	.61	19.65	2.43
	Highest				.19	.34	.46	24.07	1.65
Rating	Number					-.18	.04	3.09	.51
	Mean Quality						.85	4.25	.70
	Highest							5.45	.72
Items 7-12 (N = 16)									
Categorizing	Number		-.70	-.12	.89	-.50	-.38	3.59	.63
	Mean Quality			.51	-.66	.84	.74	16.44	2.16
	Highest				-.10	.43	.39	22.06	1.02
Rating	Number					-.43	-.26	3.80	.67
	Mean Quality						.96	4.59	.97
	Highest							5.98	.88
Items 13-18 (N = 16)									
Categorizing	Number		-.12	.21	.93	-.23	-.09	2.85	.59
	Mean Quality			.86	-.26	.92	.89	17.46	2.97
	Highest				-.03	.77	.83	21.04	2.02
Rating	Number					-.34	-.22	3.15	.60
	Mean Quality						.97	4.47	1.26
	Highest							5.61	1.26

Table 26

Correlations Between FH Scores Obtained by Categorizing and by Rating Methods (Quality Instructions)

		Categorizing			Rating			Mean	SD
		Number	Mean Quality	Highest	Number	Mean Quality	Highest		
Items 1-6 (N = 16)									
Categorizing	Number		-.37	-.15	.75	-.12	.02	1.26	.25
	Mean Quality			.97	-.51	.89	.84	19.71	4.26
	Highest				-.39	.92	.90	20.61	3.95
Rating	Number					-.32	-.16	1.21	.19
	Mean Quality						.98	5.22	1.46
	Highest							5.45	1.43
Items 7-12 (N = 17)									
Categorizing	Number		.09	.31	.77	-.13	.02	1.21	.20
	Mean Quality			.97	-.22	.76	.76	18.50	2.08
	Highest				-.01	.68	.73	19.00	2.10
Rating	Number					-.25	-.07	1.10	.17
	Mean Quality						.98	6.06	1.27
	Highest							6.19	1.26
Items 13-18 (N = 17)									
Categorizing	Number		-.03	.25	.85	.31	.35	1.42	.35
	Mean Quality			.93	-.05	.82	.77	20.00	1.88
	Highest				.21	.87	.85	21.01	1.88
Rating	Number					.23	.35	1.30	.28
	Mean Quality						.97	5.79	.95
	Highest							6.03	.96

Table 27

Reliabilities of the Six FH Scores for

GRE and Medical School Data

	GRE	Medical School (N = 58)
Best	.50	.27
Mean Quality	.62	.67
Highest	.65	.59
Number	.67	.75
Unusual	.53	.51
Unusual-High	.46	.18

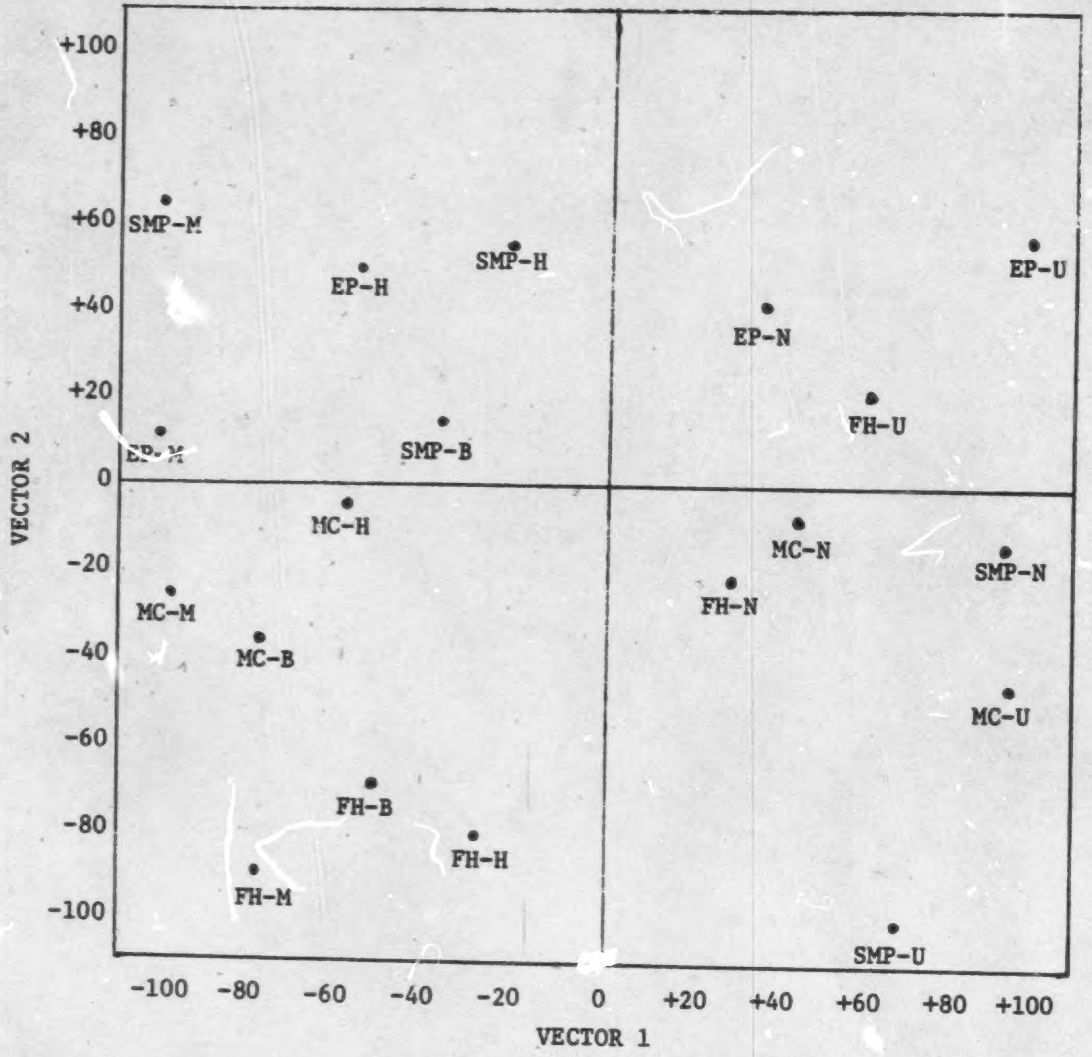
Table 28

Intercorrelations and Means of Six FH Scores for GRE and Medical School Data

		Best	Mean Quality	Highest	Number	Unusual	Unusual-High
Best	GRE		.81	.76	.14	.10	.28
	Med. Sch.		.63	.71	-.06	-.15	.14
Mean Quality	GRE			.82	-.06	-.11	.20
	Med. Sch.			.62	-.40	-.42	.12
Highest	GRE				.35	.09	.25
	Med. Sch.				.14	-.19	.15
Number	GRE					.65	.46
	Med. Sch.					.64	.40
Unusual	GRE						.67
	Med. Sch.						.50
Mean	GRE	19.67	17.88	22.09	2.45	0.61	0.24
	Med. Sch.	19.85	17.88	22.70	2.94	0.59	0.11

Figure 1

Two-Vector Smallest Space Solution[1]



---

[1] Each score is designated by the abbreviation for the test (FH, EP, SMP, MC) followed by a letter B, M, H, N, and U, respectively, designate the Best, Mean Quality, Highest Quality, Number, and Unusual scores.

Appendix A

Directions and Sample Items

FORMULATING HYPOTHESES

Directions

The problem in this test consists of a brief description of a psychological investigation, a figure or table presenting the data from the study, and a short statement of an important finding. Your task is to think of hypotheses (possible explanations) to account for the finding.

Think of the hypothesis you believe is most likely to account for the finding, and additional competing hypotheses that you think ought to be considered in interpreting the study or in planning further research. Write your hypotheses in the answer spaces. Mark the hypothesis you consider most likely to be correct by placing an X in the box at its right.

Now study the sample item and sample answers. Then write hypotheses to account for the finding shown in the test item.

Sample Problem and Answers

FORMULATING HYPOTHESES

Birth Weight and IQ

The IQ scores of 822 children aged 8 to 10 years were studied in relation to their birth weights. IQ was measured on the Wechsler Intelligence Scale for Children, while birth weights were obtained from hospital records. Results were as follows:

Relation Between Birth Weight and IQ

Birth Weight (Grams)	Number of Children	Mean IQ
1000-1499	46	84.8
1500-1999	69	88.3
2000-2499	302	91.2
2500-3000	405	95.9

Finding: Children who weighed more at birth tended to have higher IQs in middle childhood.

Suggested Hypotheses

Heavier babies came from higher SES families where mothers are more motivated regarding child care.	☒
Babies of lower weight included many premature births.	☐
mother's expectations regarding size of baby determined the level of intellectual development—i.e., large babies were encouraged to be aggressive and adventuresome.	☐
Heavier weight is due to larger and heavier brain.	☐
The same factors of the mother's diet that determine heavier birth weight also determine higher intelligence potential.	☐
	☐

Mark the hypothesis you think is best by putting an X in the box at its right.

EVALUATING PROPOSALS

Directions

Imagine that you are teaching a senior course in research design and methodology. As a class exercise, you have asked each of your students to write a brief description of a proposed experiment of his own design, following an outline you have provided. It is your plan to write criticisms of each proposal, to return the papers with your comments, and then to ask each student to revise his proposal to meet your criticisms.

One of the papers contributed by a student is shown on a later page. You are to write whatever criticisms you think are justified, whether they relate to design, methodology, analysis, or theoretical position.

Now study the sample item and sample answers. Then write your criticisms of the proposal.

Sample Item and Answers

PROPOSAL

National Prejudice

Hypothesis. Prejudice toward nationalities is the result of a lack of opportunity to interact with members of those national groups.

Procedure. Measure attitudes of American college students to fifty different nationalities, using an attitude scale. Measure on a globe the distance between the nearest borders of each country and the U. S. This gives a rough measure of opportunity to interact with members of the national group.

Analysis. Compute the correlation between the mean attitude scores for countries and distances to those countries. A significant negative correlation would verify the hypothesis.

Significance. Information about national prejudice may someday help reduce tension and war, also tests theories of prejudice.

Criticisms

There must be more and better methods for measuring opportunity to interact with foreigners!
The correlation cannot be interpreted causally.
Attitude scale responses may reflect stereotypes rather than the student's own attitude particularly for national groups about which he has no personal knowledge.
College students as a group are likely to be low on prejudice. A more representative sample might be needed to show the predicted relation.
No particular theory of prejudice is mentioned as being under Examination.

SOLVING METHODOLOGICAL PROBLEMS

Directions

The item in this test is a brief statement of a methodological problem encountered by a psychologist in planning a research investigation. Your task is to suggest ways of solving that problem. Your proposed solutions should be feasible as well as theoretically sound.

Think of the solution that you consider the best way to solve the problem, and additional solutions that you think ought to be considered in planning the study. Write your solutions in the answer spaces. Mark the solution you consider best by placing an X in the box at its right.

Now study the sample item and the sample answers. Then write solutions to the problem presented in the test item.

Sample Item and Answers

METHODOLOGICAL PROBLEM

Mathematics Attitude and Achievement

A school psychologist wishes to test the hypothesis that positive attitudes toward mathematics lead to better performance in learning mathematics. She proposes to administer an appropriate attitude questionnaire to all the ninth grade students in a school during May of the school year, and to correlate attitude scores with second-semester math grades.

A research consultant points out that, although it will be possible to discover whether there is a relation between attitude and achievement, this design will not allow any inference to be made as to the cause of the relation. For example, another possible interpretation of a positive correlation would be that students who do well in mathematics develop more favorable attitudes toward learning math.

Think of ways to discover not only whether there is a relation between attitudes toward mathematics and math performance, but also what the direction of causation is.

Suggested Solutions

Select subsamples of students whose math grades in grade 8 are equal, then look at the relation between grade 9 attitude and achievement for each subsample	☐
Measure attitudes early in the school year and math achievement at the end.	☐
Develop a program designed to improve pupil attitudes toward math and assess whether it leads to improved achievement as compared with a control group.	☒
Use a laboratory learning task, so that attitudes may be assessed before achievement has an opportunity to affect attitudes toward learning.	☐
Measure attitude and achievement both at the beginning and the end of the year. Examine all the intercorrelations of the four sets of scores.	☐
	☐

Mark the solution you think is best by putting an X in the box at its right.

MEASURING CONSTRUCTS

Directions

Measuring a psychological construct usually involves providing a set of standard conditions under which the relevant behavior can be observed. For example, the construct *spelling ability* may be measured by presenting a standard list of words, instructing the subject to mark those that are incorrectly spelled, and counting the number of mistakes.

Some constructs (for example, *extroversion*) are more often assessed indirectly, through the use of ratings or self-reports, because the appropriate behaviors are hard to elicit in a standard test situation. It is preferable, however, to use behavioral measures whenever possible. No one would consider having teachers make *ratings* of spelling ability when spelling *tests* are available.

In this test you will be given the name of a construct. Your task is to think of ways to obtain behavioral measures of the construct, that is, ways to elicit the relevant kind of behavior so that it can be observed and measured. Each method should be a reasonable possibility for use in research, whether or not it appears to be practical or efficient.

Write brief descriptions of the methods in the answer spaces. Mark the method you consider best by placing an X in the box at its right.

Now study the sample item and the sample answers. Then write ways of eliciting relevant behaviors for the construct presented in the test item.

Sample Item and Answers

CONSTRUCT

Person perception: Ability to make accurate judgments about personal characteristics of other individuals.

Suggested methods for eliciting relevant behaviors

(Please write legibly)

Have S predict the way his friends would fill out a personality inventory; then count the number of agreements with the answers given by his friends.	☐
Have S watch a film showing a discussion among six people. Then have him fill out a personality inventory on each person. Count the number of agreements with the answers actually given.	☒
Have S sort records of behavior (e.g., letters, essays, art objects) so that objects created by the same person are placed together; count the number of correct matchings.	☐
Have S predict what his friends will do in various situations; then place them in those situations and count the number of correct predictions.	☐
Have S respond in writing to some problems in the way that he thinks each of several friends would respond; then have the friends actually respond in writing. Judges then try to match predicted to actual responses.	☐
	☐

Mark the method you think is best by putting an X in the box at its right.

Appendix B

Instructions for Scoring the Tests of Scientific Thinking

Scorer's Name ______________________

Instructions
for Scoring the
Tests of Scientific Thinking

You have been assigned for scoring a problem from one of several Tests of Scientific Thinking. Attached to these instructions you should find (1) the directions for taking that test, including a sample problem and sample student responses; (2) a copy of the problem you are to score, with a blank answer sheet; (3) a specimen answer sheet as filled out by a student; (4) a specimen score sheet showing how that answer sheet was scored; (5) a blank score sheet; and (6) a list of answer categories (this is a classification of the answers frequently given to that problem).

Before doing anything else, read the test directions and study the sample problem and sample responses (Item 1 above); then respond to the problem (Item 2 above) as though you were taking the test, using the answer sheet provided. Allow yourself 8 minutes to write answers to the problem. The purpose of this is to give you a better understanding of the problem and how students might respond to it. Do not look at the student's answer sheet (Item 3) or the answer categories (Item 6) until you have written your own answers to the problem.

After responding to the problem, study the list of answer categories. See if answers essentially the same as yours in meaning are included.

The procedures to be followed in scoring an answer sheet are listed below. Please make no marks on the answer sheets as you score them.

1. Write your Scorer Number in the boxes labeled "Scorer" at the top of the score sheet. Your scorer number is 12. (The scorer who used the specimen score sheet was number 09.)

2. You are to score Problem 3 of Test A. Write these two numbers in the appropriate boxes at the top of the score sheet. (The Problem Number and Test Number shown on the specimen score sheet are 3 and A respectively.)

3. Copy the registration number of the student whose responses you are about to score in the boxes at the top of the score sheet labeled "Registration No." This four-digit number will be found (in red) on the front upper right-hand corner of each answer sheet. (The Registration No. for the Specimen Score Sheet is 2398.)

4. Locate all the Problem 3 answer sheets for the student to be scored. This will be the page labeled "MICE ULCERS AS A FUNCTION OF HOUSING CONDITION" and the reverse side of the same page.

5. The rows on the score sheet are numbered in Column 1 to correspond to the answer spaces on the answer sheet. (The answer spaces on the answer sheet are not numbered; so before you score a sheet, first number the spaces consecutively beginning on the front.) Place a plus (+) sign in each cell of Column 2 that corresponds to a space containing a scoreable response on the answer sheet.

   a. If the response in one space on the answer sheet contains two (or more) clearly different ideas, each idea should be treated as a separate response and assigned a Space Number on the score sheet. (On the Specimen Answer Sheet, the response in Space 1 includes two different ideas.) The first idea is recorded by placing a plus in Column 2 opposite the space number on the answer sheet, and the second idea is recorded by placing the same space number in the first unnumbered cell near the bottom of Column 1, together with a plus opposite that space number in Column 2. (A plus is recorded in Column 2 opposite Space 1 to indicate the first idea. The second idea is indicated by writing 1 in the first empty cell of Column 1 and placing a plus beside it in Column 2.) Treat a response as double barrelled only if it is impossible to reasonably interpret the response as a unit. If you cannot decide, consider that the response is not double barrelled. Also, if a portion of a scoreable response contains a gratuitous comment, an erroneous criticism, or other inappropriate statement (or question), disregard it and treat the response as a single unit.

b. If a single response occupies 2 (or more) answer spaces, put a plus sign only in the cell of Column 2 that corresponds to the space in which the response begins. Put a minus (-) sign in the cell (or cells) of Column 2 that correspond to spaces used for the continuation of the response. Also put a minus sign in Column 3 of each such row. (On the Specimen Answer Sheet, the first response occupies two answer spaces. A plus is placed in Column 2 opposite Space 1, and a minus is placed opposite Space 2. A minus sign is also placed in Column 3 opposite Space 2.)

c. If an answer space contains a clearly irrelevant comment (e.g., the response in Space 5 of the Specimen Answer Sheet: "By the way, I'm getting a little bored with this task!"), record a minus in the corresponding cell in Column 2, and also put a minus sign in Column 3. (On the Specimen Score Sheet, minus signs are placed in Columns 2 and 3 opposite Space 5.)

d. Make sure that the last response on the answer sheet is complete enough to permit it to be evaluated. If it is not, put minuses in Columns 2 and 3 of the appropriate row. (Space 8 on the Specimen Answer Sheet is clearly incomplete. Therefore minuses are put in Columns 2 and 3 opposite Space 8.)

e. Identify any pairs (or larger clusters) of responses that are duplicates--responses that express essentially the same idea in different words. If you find such a pair, circle the corresponding Space Numbers in Column 1 and connect the two circles with a line. (On the Specimen Score Sheet, a line connects space numbers 1 and 6 indicating that the scorer thought that the two answers were essentially the same idea.) Place a plus in Column 2 opposite the first member of the pair (or cluster), and place minuses in Column 2 and Column 3 opposite the remaining member(s) of the pair (or cluster). (On the Specimen Score Sheet, a plus is recorded in Column 2 opposite Space 1 and minuses are recorded in Columns 2 and 3 opposite Space 6.) *In all subsequent steps in scoring, treat a duplicate pair (or larger cluster) as though it were one response*.

6. The next task is to choose, for each response designated by a plus in Column 2, the category in the list of response categories that most nearly corresponds in meaning to the idea presented in that response. Place in Column 3 the number of the category corresponding to the response. (On the Specimen Score Sheet, a 15 is recorded in Column 3 opposite Answer Space 1; the scorer judged the idea in Space 1 to be similar in meaning to Category 15.)

   a. In matching responses to categories, it is not necessary that the wording be similar, only that the basic ideas are essentially similar.

   b. If the idea does not match any of the listed categories, place a capital *N* (for *None*) in Column 3 opposite the appropriate space number. (On the Specimen Score Sheet, an *N* is recorded in Column 3 opposite Space 4, indicating that the scorer did not think the idea in Space 4 was basically similar to any of the response categories.)

   c. A particular category number may be used more than once if it is appropriate for two or more responses. (On the Specimen Score Sheet, Category 12 is recorded opposite Space 3 and Space 7, indicating that the scorer thought both responses belonged to Category 12. Both responses are erroneous, yet each expresses a *different* idea.)

7. For every response designated by N in Column 3, place a number in Column 4 that represents your evaluation of the quality of the idea. Use a scale that varies from 1 (very poor) through 5 (average) to 9 (excellent). It is recommended that you leave these ratings for last; that way you will have a better feel for the various ideas that can be proposed. (On the Specimen Score Sheet, the scorer has recorded a rating of 6 in Column 4 for Response 4 to indicate that he considered it to be a little better than average.)

   a. Quality should be defined in the light of the directions to the student (see paragraph 2 of Directions). Note that these directions ask for hypotheses "most likely to account for the finding" and "additional competing hypotheses that... ought to be considered in interpreting the study or in planning further research." A statement might be a good hypothesis even though it is not in fact correct. For example, the hypothesis that data are based on an unrepresentative sample might in some instances be a reasonable one to entertain, whether or not it might prove to be correct. Ratings of the responses should, therefore, not necessarily be based on the factual accuracy of hypotheses or the assumptions on which they are based, unless you judge that a college senior majoring in psychology would be expected to know the facts.

   b. Your rating should be solely based on the quality of the idea. Try not to be influenced by such irrelevant considerations as neatness, handwriting, spelling, grammar, or elegance of writing.

c. As a general guide for rating, a 9 may be thought of as a response that you think is as good as the best response on the list of answer categories; a 1 may be thought of as a response that does not explain the finding at all, such as merely restating the finding in different words or giving an incomprehensible or ambiguous response.

8. If the student misunderstands or misreads a problem, and writes several different (nonduplicate) responses on the basis of his understanding, record plusses in Column 2 for each response and place them in Category 13. (Do not treat them as duplicates). However, if any two are truly duplicates, treat them as such by recording minuses in Columns 2 and 3 opposite the second response.

9. As you become familiar with the mechanics of scoring, you might find it preferable to score across the rows on the score sheet rather than down the columns.

10. Check your score sheet for completeness and accuracy.
    a. There should be an entry in Column 3 for every cell containing an entry in Column 2.
    b. Each plus in Column 2 should be followed either by a category number or by an N in Column 3.
    c. Each minus in Column 2 should be followed by a minus in Column 3.
    d. There should be no entries in Column 4 unless they are preceded by an N in Column 3.
    e. Do not overlook any entries near the bottom of the score sheet in the unnumbered rows.

11. Summary recording:

   a. Record the total number of plus signs in Column 2 in the boxes labeled "No. +'s Col. 2." (If the number is less than 10, put 0 in the left-hand box.)

   b. Record the sum of Column 4 in the boxes labeled "Tot. Col. 4."

   c. Record the space number of the response that the student designated as best in the box marked "Best R." (If the designated best response is a duplicate of an earlier response, record only the space number of the first one. If the designated best is a double-barrelled response, record only the idea that you think is best.)*

---

*This does not apply to the Evaluating Proposals Test.

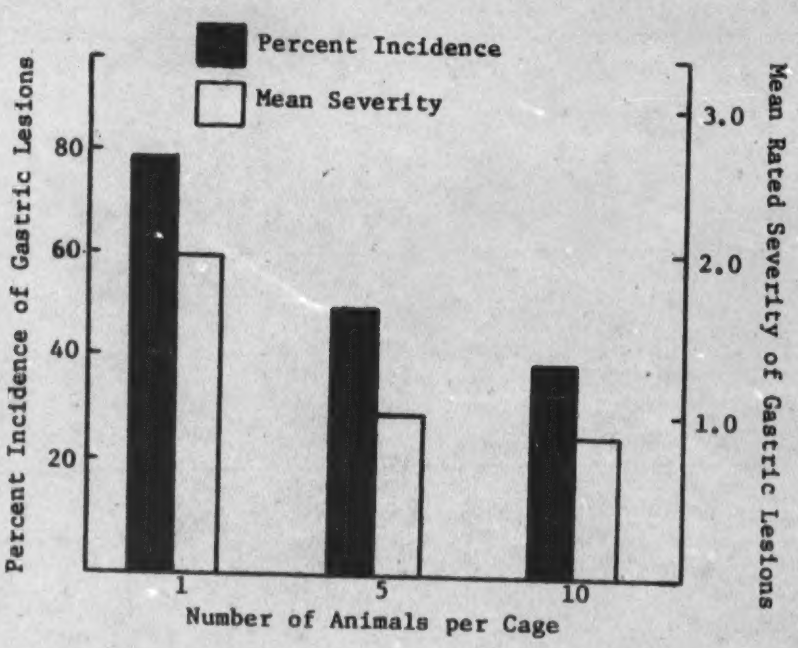
2398

FORMULATING HYPOTHESES

Mice Ulcers as a Function of Housing Condition

Thirty male mice were housed 10 to a cage (17x28x13cm) from weaning to 45 days of age. Then they were randomly assigned to different housing conditions in identical cages of the same size. They were housed either 1 per cage (N=10), 5 per cage (N=10), or 10 per cage (N=10) for seven days.

At the end of the seventh day, the mice were examined for gastric lesions (ulcers). Results are shown in the table below:



Incidence and Severity of Ulcers in Relation to Housing Conditions

Finding: The number and severity of ulcers decreased as the number of animals per cage increased.

Suggested Hypotheses

(Please write legibly)

1 The mice were accustomed to being with 9 other mice, so the change from pre-weaning to experimental treatment caused ulcers proportional to the degree of change. Furthermore, mice may be innately gregarious creatures, living together in social groups. Therefore, "solitary confinement" (the ☒

TURN THE PAGE FOR MORE ANSWER SPACES

Suggested Hypotheses

2	one mouse per cage condition) may be a stressful condition, leading to ulcers.	☐
3	Judging from the number of ulcers, I think the results may be due to something wrong in the experimental design.	☐
4	Single mice suffer from the lack of physical warmth probided by the presence of other mice when living together.	☐
5	By the way, I'm getting a little bored with this task!	☐
6	The contrast between what the mice were used to & their new conditions caused the ulcers.	☐
7	Perhaps mice were housed in different-sized cages, which led to a difference in ulcer-formation.	☐
8	Perhaps single mice didn't get enough	☐

Mark the hypothesis you think is best by putting an X in the box at its right.

SCORE SHEET

Scorer 0 9

Problem 3

Test A

Registration No. 2 3 9 8

1 Space No.	2 Response	3 Category No.	4 Rating
①	+	15	
2	—	—	
3	+	12	
4	+	N	6
5	—	—	
⑥	—	—	
7	+	12	
8	—	—	
1	+	16	

No. +'s Col. 2 0 5

Tot. Col. 4 0 6

"Best" R 1

FH ANSWER CATEGORIES

General

1. There were too few cases to draw conclusions.
2. There was bias (unspecified) in assigning Ss to treatments.
3. The sample was not typical (representative) of the population (in ways unspecified).
4. Errors (unspecified) in the design or conduct of the study could account for the finding.
5. The experimenter, knowing the purpose of the experiment, was biased in his treatment of the groups.
6. The experimenter (observer, evaluator), knowing the purpose of the experiment, was biased in his assessment of the results.
7. The measurement procedure (instrument, test) was inadequate (not valid, unreliable).
8. The statistical method was inappropriate (inadequate).
9. The results are not statistically significant.
10. [The response is incomprehensible (illegible, ambiguous, vague).]
11. [The response is essentially a restatement of the finding.]
12. [The response is erroneous or is an erroneous criticism of the experimental design or procedure.]
13. [The examinee apparently misread or misunderstood the problem.]
14. [The response does not explain the finding, although it may explain something else, or it may be a gratuitous comment or observation regarding the study.]

FH ANSWER CATEGORIES

FH-3. Mice Ulcers as a Function of Housing Condition

15. The contrast (change) between initial housing and new housing was associated with an increase in ulcers.

16. Since mice are social animals, separation from the group produced stress (anxiety, fear) which led to ulcers (dominance hierarchies were disrupted).

17. Separation from the group caused loneliness, boredom, loss of appetite, reduced activity, which led to ulcers.

18. Separation from the group caused sexual frustration which led to ulcers.

19. Mice living in larger groups had other mice upon which to release tension and stress; in single mice, this stress was inwardly directed.

20. The change in housing condition occurred at a critical period in the lives of the mice, when they need to be with other mice or when major neurological development takes place.

21. As the number of mice decreased and the available food increased, eating habits changed, which produced more ulcers.

22. Lack of social grooming (e.g., licking, lice removal) in the isolated mice produced stress which led to ulcers.

23. Mice in larger groups had less room for movement; so they became less active (more relaxed), hence had fewer ulcers.

24. Ulcers were caused by excessive space for single mice.

25. Single mice have difficulty coping by themselves, whereas several mice work together to survive.

Appendix C

Statistical Procedures

Appendix C

Statistical Procedures

A. Test Reliability

The most frequently used test reliability coefficient is coefficient alpha (Cronbach, 1951), which is defined:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_T}\right)$$, where n = the number of items

$\sum_i V_i$ = sum of variances of items

V_T = variance of total score on the test

The variance of the total score may be computed from the item variance-covariance matrix, since:

$$V_T = \sum_i V_i + \sum_i \sum_{j \neq i} C_{ij} \ ;$$

i.e., the sum of the item variances plus the double sum of item covariances. The reliability sample in the present study provides all the necessary item variances and covariances for this computation. Alpha provides a lower bound internal consistency estimate.

However, a superior lower bound estimate is available; this is the coefficient labeled by Guttman (1945) "Lambda two" and by Koutsopoulos (1964) "Lambda three squared." The formula is:

$$\lambda_3^2 = 1 - \frac{\sum_i V_i}{V_T} + \frac{\sqrt{\frac{n}{n-1} \sum_i \sum_{j \neq i} C_{ij}^2}}{V_T}$$

Koutsopoulos provides a proof that lambda is at least as large as alpha; in general it provides a higher estimate, and therefore underestimates "true" reliability by a lesser amount. Apparently, the more general use of alpha, despite the superiority of this second coefficient, is attributable to computational convenience. In the usual situation in which total score variance is available, alpha can be computed without generating the matrix of item covariances, while lambda requires this extra step. In the present investigation, lambda has been used; it provides the lower bound entries in Table 6. (Alphas were also computed for the present data; they average .05 lower than lambda, with the differences ranging from .01 to .15 for individual coefficients.)

Table 6 also contains estimates which we have referred to as "upper bound" or "parallel forms test-retest" reliability estimates. These are estimates of the test-retest correlation to be expected if two hypothetical forms of a six-item test which were parallel in all respects were given to the same group of individuals. Following Cronbach (1951), the largest split-half correlation, corrected for length, provides the estimate.

Each of the ten possible split-half correlations for a test of six items was computed by the formula:

$$r_{AB} = \frac{\sum_{A}\sum_{B} C_{AB}}{\sqrt{V_A}\,\sqrt{V_B}}$$

where A is a set composed of one-half of the test items and B is a set composed of the remaining items; the numerator is the sum of all covariance terms involving one item from set A and one from set B; and the denominator is the product of the square roots of the total variances of scores on sets A and B. These total variances, in turn, are obtained by:

$$V_A = \sum_i V_{A_i} + \sum_i \sum_{j \neq i} C_{A_{ij}} \quad ;$$

that is, the sum of the variances of items in set A plus the double sum of covariances of items in set A.

The largest of the ten split-half correlations was then corrected by the Spearman-Brown prophecy formula,

$$r_2 = \frac{2r}{1 + r} \quad ,$$

to provide an estimate for a test of six items.

B. Score Means and Standard Deviations

Means and standard deviations of scores for six-item tests are presented in Table 7, based on data from the reliability sample, and in Table 9, based on data from the intercorrelation sample. The means in Table 7 are the means over the six items making up a test, without weighting for the (relatively small) differences in numbers of candidates contributing data for a particular item. The means in Table 9 are similarly the unweighted means over six items times three contexts in which an item was given (for example, items in FH were given sometimes along with EP, sometimes with SMP, and sometimes with MC). Unweighted means are appropriate since the interest here is in estimating the results which would have been obtained had complete six-item tests been given to individuals.

Score standard deviations are also presented on a per-item basis. To obtain the values given in Table 7, estimated variances for total scores were calculated from the variance-covariance matrices used in deriving reliability estimates, by the formula:

$$V_T = \sum_i V_i + \sum_i \sum_{j \neq i} C_{ij} \quad .$$

The standard deviation of the total score is then the square root of the variance, and that of the mean score is one-sixth that of the sum.

Estimation of standard deviations for the intercorrelation sample data requires the use of information from the reliability sample, since it is only in the latter group that the necessary within-test item covariances are to be found. As discussed in the text, the similarity of corresponding means and of corresponding item variances across contexts supports the use of this procedure for the quality scores. For the count scores, however, a better approximation can be made by treating an item given in a two-item context as constituting a longer test than the same item given in a three-item context. The analogy is somewhat plausible, in that a likely explanation for the differences is that candidates taking an item in a two-item context had more time in which to generate responses to that item. In any case, the assumption of differing length provides a basis for adjusting item covariance estimates from the reliability sample in appropriate proportion to the differences across contexts in item variances.

Following Gulliksen (1950):

$$M_k = kM_1 \quad ;$$

that is, the mean of a test of length _k_ is equal to _k_ times the mean of a unit-length test. Then:

$$s_k = s_1 \sqrt{k + k(k-1)r_{11}} \quad ;$$

that is, the standard deviation of a test of length k may be expressed in terms of the standard deviation of the unit-length test, the lengthening factor k, and the reliability of the unit-length test.

The lengthening factor k was estimated for each count score for each test by taking the ratio of corresponding means from Tables 7 and 9. k ranged from 1.05 to 1.28 for the various test by score combinations, with an average value of 1.15. Then, for each combination, k and the appropriate parallel-forms reliability estimate from Table 6 were used to derive a corrected estimate of total variance for the score. These estimates were used in computing the standard deviations for count scores reported in Table 9.

C. Correlations of Scores Within Each Test

The correlation between two scores within a test is obtained using data from the reliability sample, and is computed:

$$r_{AB} = \frac{\sum_i \sum_j C_{ij}}{\sqrt{V_A} \sqrt{V_B}}$$

where A is one score on a test, B is another; the numerator is the sum of the (36) covariance terms involving one score A on item i and the second score B on item j, i and j = 1 - 6; and each term in the denominator is a total score variance estimated as described in earlier sections of this appendix.

The covariance terms can be separated into two subsets. Those for $\underline{i} = \underline{j}$ are terms expressing the relation of two scores derived from data on a single item; since the same performance provides the basis for both scores, the relation is inflated by experimental interdependencies. Those for $\underline{i} \neq \underline{j}$ are terms expressing the relation of one score from one item with another score from a second item; here there is no such interdependency. An estimate of the intercorrelation of two scores free of experimental interdependencies may be made by assuming that the covariance terms for $\underline{i} = \underline{j}$ would, on the average, equal those for $\underline{i} \neq \underline{j}$ if different performance were used throughout as the basis for obtaining each score on a six-item test. This assumption is equivalent to replacing the numerator in the expression above by:

$$(36/30) \sum_i \sum_j C_{ij}, \; i = 1\text{-}6, \; j = 1\text{-}6, \; i \neq j \; .$$

Finally, estimated true score correlations between two scores from a test are obtained by the formula:

$$r_{TT} = \frac{r_{AB}}{\sqrt{r_{AA}}\,\sqrt{r_{BB}}} \quad ;$$

that is, the correlation which would have been obtained had each of the abilities represented by one of the scores been measured without error is obtained by dividing the obtained coefficient by the product of the square roots of the reliabilities of the scores. For these calculations, the more conservative parallel-forms reliability coefficients (i.e., those leading to lower estimates) were employed. The coefficients which were stepped up were those from Table 11, from which the effects of experimental interdependencies were removed, since the interest is in estimating the true relation between the abilities underlying the obtained scores rather than between the scores themselves.

D. Correlations of Scores from Different Tests

The correlation between two scores from different tests is given by:

$$R_{AB} = \frac{\sum_i \sum_j C_{ij}}{\sqrt{V_A}\,\sqrt{V_B}}$$

where the i's are the six items from one test and the j's are the six items from the second; and where each term in the denominator is a total score variance. For the quality scores the variances employed were those used in all the computations described above, while for the count scores they were the total variances adjusted for differences in test "length" as described in Section B of this appendix.

These correlations were also computed using an alternate procedure for estimating total variances of scores in the intercorrelation sample. Here, an item i from one of the tests was considered to constitute a unit-length test, and the problem was seen as that of estimating the variance of a test lengthened to six items. For each score, within the data for candidates given items from each pair of tests, the standard deviation of the unit length test was taken to be the square root of the mean item variance for the six items from that test. By Gulliksen's (1950) formula:

$$s_k = s_1 \sqrt{k + k(k-1)r_{11}} \quad ,$$

where s_k is the standard deviation of the lengthened test, s_1 is that of the unit-length test, k is the factor by which the test is lengthened (and, by assumption, equals 6), and r_{11} is the reliability of the unit-length test. All the terms required in this computation are directly available from the performance of the relevant subgroup from the intercorrelation sample, except for the reliability estimate. The key assumption in this procedure is that the reliability of a score remains constant across testing contexts, so that results from the reliability sample can provide the necessary coefficients. Using coefficient alpha as the reliability estimate, the Kuder-Richardson formula 20 taken in reverse allows estimation of the reliability of a one-item test from that of the six-item test; i.e., if:

$$r_n = \frac{nr_1}{1 + (n-1)r_1}$$

then:

$$r_1 = \frac{r_n}{n(1-r_n) + r_n}$$

where r_n is the reliability of the longer (6-item) test, r_1 is that of a one-item test, and $n = 6$.

As indicated in the text, intertest correlations computed using the two different methods for estimating test total variances which have been described were similar both in absolute and relative magnitudes. The two approaches differ substantially in the way in which information from the reliability sample is employed in estimating variances for the intercorrelation sample data; thus, these results justify some confidence in the pattern of relations which is obtained.

A further adjustment procedure which was considered should also be mentioned. This method requires the assumption that the correlation between two items from a test is constant from the reliability to the intercorrelation samples. Then, each within-test item covariance term required can be obtained by

$$C_{ij} = r_{ij}\, s_i\, s_j \quad ,$$

where s_i and s_j are the item standard deviations from the relevant intercorrelation sample matrix and r_{ij} is the correlation between items $\underline{i}$ and $\underline{j}$ in the reliability sample data. However, on the further assumptions that each r_{ij} is an estimate of a common inter-item correlation (i.e., that the test is homogeneous) and that each s_i is an estimate of a common item standard deviation, this assumption proves equivalent to assuming that the ratio of total test variance from the reliability sample to that from the intercorrelation sample is equal to the ratio of the corresponding mean item variances. This, in turn, is the result which would be obtained following the "lengthening" analogy with estimated test reliabilities of 1.00. This assumption would clearly be unreasonable.

E. Correlations with GRE Scores

The correlation between a score on test A and one of the GRE scores B is derived from reliability sample data and is given by:

$$R_{AB} = \frac{\sum_i C_{ij}}{\sqrt{V_A}\,\sqrt{V_B}}$$